Butlletí de la Societat Catalana de Matemàtiques
Vol. 24, núm. 1, 2009. Pàg. 5–12. DOI: 10.2436/20.2002.01.18

La llei de Benford*

Elise Janvresse, Thierry de la Rue

Resum La llei de Benford reflecteix una irregularitat inesperada en la distribució dels dígits de dades aleatòries. En aquest article es descriu el fenomen i se'n discuteixen possibles explicacions.

Paraules clau: llei de Benford, distribució de dígits.

Classificació MSC2010: 62-01, 62D05.

Els nombres que es troben a la vida quotidiana (temperatures, dates, preus, valors a la borsa, etc.) obeeixen una llei força inesperada. Podem classificar aquests nombres segons la primera xifra significativa, que estarà compresa entre 1 i 9; no és mai 0 i no es té en compte ni el signe ni el lloc de la coma. Així, els nombres $0,021$, $25,6$ o -2 tenen com a primera xifra significativa el 2. Si es consideren prou nombres, d'orígens diversos, es manifesta una tendència sorprenent: es troben molts més nombres començant per 1, 2 o 3, que per 7, 8 o 9.

El primer de comentar aquest fenomen, que el va descriure l'any 1881 en un article de l'*American Journal of Mathematics* ([7]), va ser l'astrònom Simon Newcomb. Es va sorprendre que les taules de logaritmes es desenquadernessin més sovint per les primeres pàgines que per les darreres. El logaritme és una funció matemàtica gràcies a la qual es poden transformar les multiplicacions i les divisions (operacions molt complicades sense calculadora) en sumes i restes; per tant, es pot imaginar fàcilment l'ús intensiu d'aquestes taules pels científics de l'època. L'explicació proposada per Newcomb fou que les taules s'utilitzaven més pel principi que pel final perquè els usuaris trobaven més sovint nombres que començaven per 1 o 2 que per 8 o 9. Llavors, va proposar la fórmula següent, en la qual la probabilitat d'aparició de cadascuna de les xifres possibles es descriu precisament amb ajuda de la funció logaritme (en

* Traducció de Frederic Utzet. Aquest treball va resultar finalista en el concurs d'articles de divulgació matemàtica convocat per l'European Mathematical Society el 2006. Agraïm als autors l'autorització que amablement ens han donat per a publicar aquesta traducció al català.

base 10): per $i \in \{1, \ldots, 9\}$, la primera xifra significativa és i amb probabilitat

$$P(i) = \log_{10}\left(1 + \frac{1}{i}\right).$$

Cal notar que es tracta d'una probabilitat:

$$P(1) + P(2) + \cdots + P(9) = \log_{10} 10 = 1.$$

El físic Frank Benford, que no coneixia l'article de Newcomb, va fer la mateixa descoberta cinquanta-set anys més tard, també notant l'ús irregular de les taules de logaritmes. Ell també va publicar un article sobre el tema ([1]), en el qual proposava la mateixa fórmula, obtinguda empíricament, per a descriure la freqüència segons la qual es distribueix la primera xifra significativa. Atès que l'article de Benford va tenir més ressò que el de Newcomb, aquesta fórmula porta des de llavors el nom de *llei de Benford*. A la taula1 es troben els valors numèrics aproximats proporcionats per aquesta llei.

$P(1)$	$P(2)$	$P(3)$	$P(4)$	$P(5)$	$P(6)$	$P(7)$	$P(8)$	$P(9)$
30,1%	17,6%	12,5 %	9,7 %	7,9%	6,7%	5,8%	5,1%	4,6%

TAULA 1: Valors numèrics aproximats de la llei de Benford. Les xifres 1 i 2 tenen quasi la meitat de tots els nombres!

Hem tractat de comprovar la validesa de la llei de Benford utilitzant nombres agafats a l'atzar en diaris i catàlegs: per a cada categoria —nombres que apareixen en els articles, valors de la borsa, preus en euros, els mateixos preus convertits en francs— hem buscat algunes desenes de nombres. Les figures 1 i 2 comparen la llei de Benford amb les distribucions de les primeres xifres trobades. La semblança és en principi una mica feble (vegeu la figura 1), però quan s'acumulen totes les dades (460 en total) és sorprenent (vegeu la figura 2).

Benford va comprovar la seva llei amb un nombre considerable d'observacions —tan gran com el temps i l'energia humanament disponible li permeten— d'orígens diversos: en total va recollir exactament 20.299 nombres que procedien des de resultats de beisbol fins a anotacions d'hidrologia. Actualment la informàtica permet contrastar molt ràpidament la llei de Benford en conjunts de dades molt més grans. Tal com suggereix l'economista americà Marc Nigrini (vegeu la darrera part de l'article) a la seva pàgina web, [1] hem considerat la mida de 56.544 fitxers que hi havia en el disc dur d'un dels nostres ordinadors (vegeu la figura 3). Curiosament, les xifres 1 i 4 apareixen clarament molt més sovint que el que preveu la llei de Benford. Una observació més detallada de les mides dels fitxers mostra que els nombres 16.384 i 4.096 (respectivament, 2^{16} i 2^{14}) s'obtenen milers de vegades cadascun. Això és a causa que són les mides que el sistema reserva automàticament per als repositoris. Si no es tenen en compte aquests fitxers, la distribució obtinguda és extremadament semblant a la llei de Benford!

1 http://www.nigrini.com/benford's_law.htm

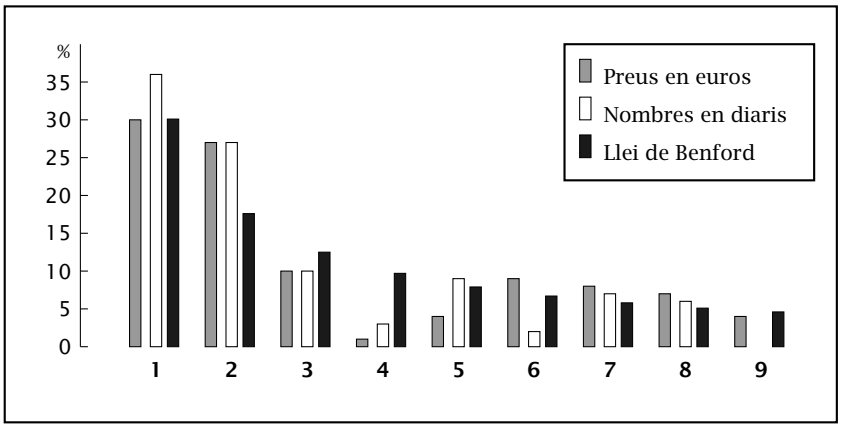

FIGURA 1: Comparació de la llei de Benford amb les freqüències relatives observades de la primera xifra significativa de nombres agafats a l'atzar en diaris i catàlegs.

Per què surt més l'1 que el 9?

Atès que és possible escriure tants nombres que comencin per 1 com per 9, per què se'n troben més sovint dels primers que dels segons? Prenem un exemple senzill de la data d'aniversari (mes i dia del mes) d'un nombre gran de persones. Podem suposar que cadascun dels mesos, representat per un nombre entre 1 i 12, apareixerà una vegada de cada 12. Aleshores, 1 serà la primera xifra representativa en un terç dels casos, i cadascuna de les altres xifres apareixerà amb una freqüencia $1/12$. Si tenim en compte el dia i el mes, veiem que la distribució de xifres s'assembla una mica a la donada per la llei de Benford: privilegia clarament 1 i 2 enfront de les altres xifres. La conclusió és la següent: és veritat que hi ha tants nombres començant per 1 com per 9 entre 1 i 999, o entre 1 i 9.999, i s'observa una freqüència uniforme si es miren, per exemple, les matrícules dels cotxes amb els quals ens creuem al carrer. Però això no és veritat entre 1 i 12, o entre 1 i 31, etc. De fet, és fals en tant que el màxim no és de la forma $10^n - 1$. Segons la natura de les dades observades, el màxim varia, però l'1 és més freqüent que el 9.

La primera idea que ve al cap quan es vol demostrar rigorosament la llei de Benford és intentar provar-la a $\mathbb{N}$, mirant primer de tot el conjunt D_1 dels nombres naturals que comencen per 1. Dissortadament, D_1 no té una densitat natural entre els nombres naturals: la quantitat

$$\alpha_1(n) = \frac{1}{n}\,|D_1 \cap \{1, 2, \ldots, n\}| = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{D_1}(k) \tag{1}$$

no para d'oscil·lar entre $1/9$ i $5/9$ quan n varia. Per tant, $\alpha_1(n)$ no té límit quan n tendeix a ∞. El mateix passa quan s'estudien les altres primeres xifres significatives possibles.

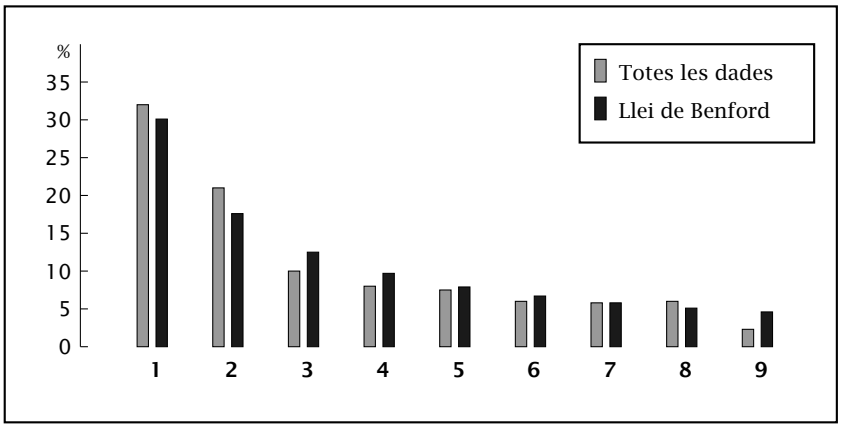

FIGURA 2: En aquesta gràfica s'han reagrupat els 460 nombres obtinguts a l'atzar; ens aproximem a la llei de Benford...

Flehinger ([2]) proposa definir una *densitat generalitzada* obtinguda iterant el procés de calcular la mitjana (mitjana de Cesaro): per a $t > 1$, posem

$$\alpha_t(n) = \frac{1}{n} \sum_{k=1}^{n} \alpha_{t-1}(k).$$

Flehinger, al seu article, prova que l'amplada de les oscil·lacions de les funcions $\alpha_t(n)$ disminueix i que el procediment convergeix en el sentit següent:

$$\lim_t \liminf_n \alpha_t(n) = \lim_t \limsup_n \alpha_t(n) = \log_{10} 2.$$

Es troba aleshores la probabilitat esperada de tenir un nombre que comença per 1, $\log_{10}(1 + 1/1) = \log_{10} 2$. Aquest resultat és igualment correcte per a les altres primeres xifres significatives possibles. També es troba una demostració del resultat de Flehinger a [6].

Llei de Benford general per a totes les xifres significatives

En la seva formulació més general, la llei de Benford no descriu només la distribució de la primera xifra significativa, sinó també la de la segona, la tercera, i totes les següents. Per enunciar de manera simple aquesta llei de Benford general, és còmode utilitzar la *mantissa* d'un nombre $x > 0$. Es pot interpretar la mantissa de x com el nombre que queda quan s'oblida el lloc de la coma a l'escriptura de x en base 10. Més formalment, donat un nombre real $x > 0$, es defineix la mantissa[2] de x com l'únic real $\mathcal{M}(x) \in [1, 10[$ tal que es pot trobar

2 De fet, normalment la mantissa es defineix com un nombre entre 1/10 i 1, però amb relació a la llei de Benford veurem que és més pràctic considerar-la a $[1, 10[$.

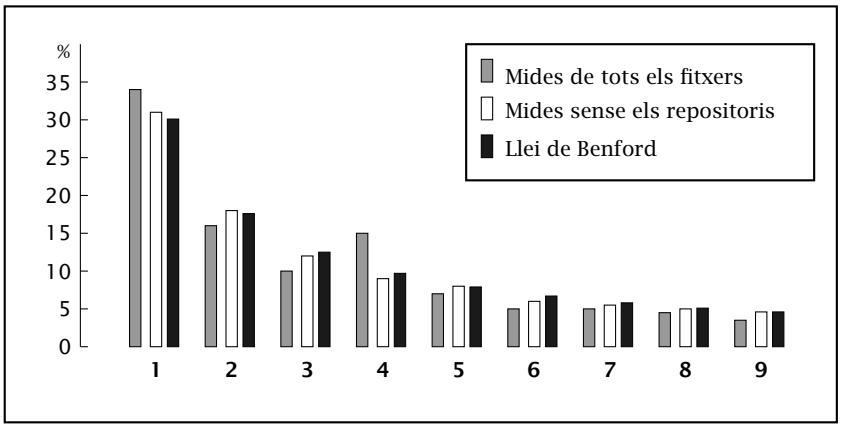

FIGURA 3: Distribució de la primera xifra significativa de la mida de 56.544 fitxers; si no es tenen en compte els fitxers repositoris, la llei de Benford és sorprenentment ben respectada.

un enter k que compleix

$$x = \mathcal{M}(x)10^k.$$

Per exemple, $\mathcal{M}(2002) = \mathcal{M}(0,02002) = 2,002$. Així, la mantissa de x permet conèixer totes les xifres significatives de x en base 10, però no diu res sobre l'ordre de magnitud de x.

La llei de Benford general descriu precisament la distribució de la mantissa: segons aquesta llei, la proporció de nombres x tals que $\mathcal{M}(x) \in [a, b[$, on $1 \leq a < b \leq 10$ és

$$P_{Benford}(\mathcal{M}(x) \in [a, b[) = \log_{10} b - \log_{10} a. \tag{2}$$

Remarquem que $\log_{10} 10 - \log_{10} 1 = 1$, així el 100 % dels nombres té la mantissa entre 1 i 10, la qual cosa és un bon senyal. A més, per a tot $i \in \{1, \ldots, 9\}$, dir que la primera xifra significativa de x és i es tradueix en termes de la mantissa per

$$\mathcal{M}(x) \in [i, i + 1[,$$

que segons (2) porta a la probabilitat

$$\log_{10}(i + 1) - \log_{10} i = \log_{10}\left(\frac{i+1}{i}\right) = \log_{10}\left(1 + \frac{1}{i}\right).$$

Així, retrobem la probabilitat anunciada anteriorment per a la primera xifra significativa.

De manera força sorprenent, com remarca Ted Hill a [4], la formulació general de la llei de Benford implica que les distribucions de les xifres significatives successives no són independents! Per exemple, entre els nombres la primera

xifra dels quals és 1, la proporció dels que tenen com a segona xifra 0 és

$$\frac{\log_{10} 1,1 - \log_{10} 1}{\log_{10} 2 - \log_{10} 1} \simeq 13,7\ \%,$$

mentre que entre aquells en què la primera xifra és 9, la xifra 0 apareix en segona posició amb probabilitat

$$\frac{\log_{10} 9,1 - \log_{10} 9}{\log_{10} 10 - \log_{10} 9} \simeq 10,5\%.$$

Invariància per canvi d'escala

Tot i que la llei logarítmica de Benford donada per (2) és *a priori* sorprenent, la forta adequació a les dades experimentals amb aquesta formulació ha empès els matemàtics a trobar-ne justificacions *naturals*. Un panorama força ample d'aquests treballs es troba a l'article de Ralph A. Raimi ([9]), i altres arguments han estat donats posteriorment per Ted Hill ([3, 5]). Entre totes aquestes explicacions hem escollit presentar-ne una de les que semblen més convincents. Va ser proposada per primer cop per Roger S. Pinkham el 1961 ([8]) i va ser represa als articles de Raimi i Hill.

Les lleis matemàtiques no han de tenir fronteres!

Suposem que existeix una llei matemàtica que prediu quina és la proporció de nombres trobats a la vida quotidiana que tenen la mantissa entre a i b. El mínim que podríem demanar és que la llei fos la mateixa per a tot el món! Hauria de ser idèntica en un país on les distàncies s'expressin en metres i en un país que utilitzin el sistema anglosaxó. Hauria de ser igual a la zona euro, i a la zona dòlar, o yen o lliura esterlina. Dit d'una altra manera, aquesta llei hauria de ser independent de les unitats amb les quals s'expressen les quantitats.

Com traduir aquesta propietat matemàticament? Canviar la unitat amb la qual s'expressa una quantitat x vol dir multiplicar x per un factor de conversió α que és la proporció entre les dues unitats (per exemple, $\alpha = 1/166,386$ quan es passa de pessetes a euros). La llei que descriu la distribució de les mantisses ha de ser invariant per multiplicació per un factor α qualsevol. Això es pot escriure de la manera següent: per a qualsevol $\alpha > 0$ i $1 \leq a < b \leq 10$,

$$P(\mathcal{M}(x) \in [a,b[) = P(\mathcal{M}(\alpha x) \in [a,b[). \tag{3}$$

La clau per a entendre les conseqüències d'aquesta equació és treure logaritmes, de cara a transformar la multiplicació per α en una suma. Atès que es treballa amb nombres escrits en base 10, s'imposa la utilització de logaritmes decimals. Llavors, per definició de mantissa,

$$log_{10}(\mathcal{M}(x)) = \{\log_{10} x\},$$

on $\{y\}$ designa la part fraccionària de y, és a dir, l'únic nombre a $[0,1[$ tal que $y - \{y\}$ sigui enter. El terme a l'esquerra de l'equació d'invariància (3) s'escriurà

$$P(\mathcal{M}(x) \in [a,b[) = P(\{log_{10}x\} \in [\log_{10} a, \log_{10} b[),$$

i el de la dreta es transforma en

$$P(\mathcal{M}(\alpha x) \in [a,b[) = P(\{log_{10}x + log_{10}\alpha\} \in [\log_{10} a, \log_{10} b[).$$

Com que la part fraccionària d'una suma és igual mòdul 1 a la suma de les parts fraccionàries, podem escriure

$$P(\mathcal{M}(\alpha x) \in [a,b[) = P(\{log_{10}x\} \in [\log_{10} a - log_{10}\alpha, \log_{10} b - \log_{10}\alpha[),$$

on $\log_{10} a - log_{10}\alpha$ i $\log_{10} b - log_{10}\alpha$ s'han d'entendre mòdul 1. Fem un canvi de variables posant $s = \log_{10} a$, $t = \log_{10} b$ i $u = -\log_{10}\alpha$. L'equació d'invariància (3) esdevé: per a qualsevol $0 \le s < t \le 1$ i qualsevol real u,

$$P(\{\log_{10} x\} \in [s,t[) = P(\{\log_{10} x\} \in [s+u, t+u[), \tag{4}$$

on, com abans, $s+u$ i $t+u$ estan definits mòdul 1.

Així, la proporció dels x tals que $\{\log_{10} x\} \in [s,t[$ donada per la llei només pot dependre de la llargada de l'interval $[s,t[$. En particular, si $[s,t[$ és de la forma $[k/n, (k+1)/n[$ amb k i n enters, aquesta proporció ha de valer $1/n$, ja que els n intervals disjunts recobreixen $[0,1[$. Aleshores es dedueix, primer per a s i t racionals, i després mitjançant un pas al límit per a s i t reals que

$$P(\{log_{10}x\} \in [s,t[) = t - s.$$

Retornant a la mantissa de x i a les variables a i b, això dóna

$$P(\mathcal{M}(x) \in [a,b[) = \log_{10} b - \log_{10} a.$$

Per tant, la distribució P és la llei de Benford!

Per què serveix tot això?

Com moltes descobertes matemàtiques, la llei de Benford ha estat molt temps una curiositat sense cap aplicació pràctica, fins que als anys noranta del segle passat l'economista americà Mark Nigrini va suggerir la utilització de tests basats en aquesta llei per a la detecció de dades falsificades (vegeu el web citat anteriorment). Nigrini va demostrar que un examen acurat dels nombres que sortien a la comptabilitat d'un negoci podia permetre a un comptable expert detectar possibles fraus. En efecte, l'experiència mostra que les dades autèntiques han de seguir la llei de Benford. Al contrari, qui «inventa» nombres té tendència a sobreestimar l'aparició de 5 i 6. A la pràctica es fan servir tests més fins que fan intervenir la distribució de les dues primeres xifres significatives. Aquests tests han permès trobar falsificacions a la comptabilitat de set empreses de Nova York. Conclusió: tant als inspectors com als estafadors els convé conèixer la llei de Benford!

Referències

[1] Benford, F. «The law of anomalous numbers». *Proc. Amer. Phil. Soc.*, 78 (1938), 551–572.

[2] Flehinger, B. J. «On the probability that a random integrer has initial digit *a*». *Amer. Math. Monthly*, 73 (1966), 1056–1061.

[3] Hill, T. «Base-invariance implies Brenford's law». *Proc. Amer. Math. Soc.*, 123 (1995), 887–895.

[4] Hill, T. «The significant digit phenomenon». *Amer. Math. Monthly*, 102 (1995), 322–327.

[5] Hill, T. «A statistica derivation of the significant-digit law». *Stat. Science*, 10 (1996), 354–363.

[6] Knuth, D. *The art of computer programing*. Reading, Massachusetts: Addison-Wesley Publishing Company, 1997. Vol. 2.

[7] Newcomb, S. «Note on the frequency of use of the different digits in natural numbers». *Amer. J. Math.*, 4 (1881), 39–40.

[8] Pinkham, R. S. «On the distribution of first significant digits». *Ann. Math. Statist.*, 32 (1961), 1223–1230.

[9] Raimi, R. A. «The first digit problem». *Amer. Math. Monthly*, 85 (1976), 521–538.

Laboratoire de Maths Raphaël Salem
UMR 6085 CNRS
Faculté des Sciences
Université de Rouen
Saint Étienne du Rouvray, França
Elise.Janvresse@univ-rouen.fr
delaRue.Thierry@univ-rouen.fr