La legge di Benford

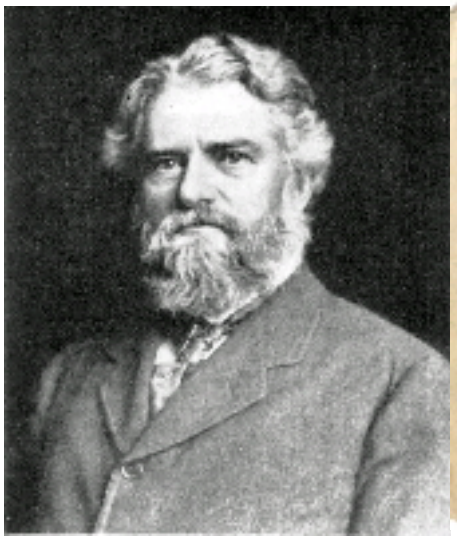
L. Brandolini e G. Travaglini

Sunto. *Guardando le pagine contenenti le tavole dei logaritmi (la carta delle pagine, non solo quello che vi era stampato!) un astronomo inglese di fine 800 si accorse di una apparente stranezza nella distribuzione delle prime cifre dei numeri che appaiono in moltissimi fenomeni. Oggi questa stranezza è diventata una teoria non ancora ben compresa, ma ampiamente utilizzata in statistica, informatica e nell'investigazione delle frodi, ad esempio quelle fiscali.*

1 Introduzione

In un breve articolo pubblicato nel 1881 sull'*American Journal of Mathematics* ([7]) l'astronomo Simon Newcomb scriveva queste righe, nate dall'avere osservato da un diverso punto di vista un oggetto allora di uso comune per scienziati e tecnici: le tavole dei logaritmi.

> *Che le dieci cifre non appaiono con uguale frequenza deve essere evidente a chiunque faccia molto uso delle tavole dei logaritmi, e noti che le prime pagine sono più consumate delle ultime. La prima cifra significativa è* 1 *più spesso che un'altra cifra, e la frequenza diminuisce fino al* 9 *... La legge della probabilità dell'apparire dei numeri è tale che tutte le mantisse dei loro logaritmi sono equiprobabili.*



S. Newcomb

Vecchie tavole di logaritmi

Perché una volta si usavano le tavole dei logaritmi?
Fino all'avvento delle macchine calcolatrici i logaritmi (o meglio le tavole logaritmiche e il regolo calcolatore) sono stati uno strumento utile ed estremamente diffuso per lo svolgimento di calcoli complicati. Per moltiplicare due numeri positivi era sufficiente passare ai loro logaritmi, sommarli e poi tornare indietro; con il vantaggio che la somma è un'operazione molto più agevole del prodotto. Oppure, per esempio, il calcolo di una radice n-esima di un numero positivo era ridotto alla divisione per n del suo logaritmo, e anche qui la divisione è più semplice della radice n-esima.

La *prima cifra significativa* di cui parlava Newcomb è la prima cifra diversa da 0 presente nello sviluppo decimale del numero. Per esempio, la prima cifra significativa di $3,14159265\ldots$ è 3, la prima cifra significativa di 2012 è 2, la prima cifra significativa di $1/2012 = 0,000497017893\ldots$ è 4.

Cerchiamo ora di interpretare l'ultima affermazione: *le mantisse dei loro logaritmi sono equiprobabili.*

Indichiamo con $[x]$ la parte intera di un numero reale x (cioè il più grande intero che non supera x) e con $\langle x\rangle = x - [x]$ la sua parte frazionaria (o mantissa). Quindi, ad esempio,

$$\begin{array}{ll} [\pi] = 3, & \langle\pi\rangle = 0,14159265\ldots \\ [-1,25] = -2, & \langle -1,25\rangle = 0,75. \end{array}$$

Qualsiasi numero reale positivo v può essere scritto nella forma

$$v = 10^M w,$$

con M intero (positivo, negativo o nullo) e $1 \leq w < 10$. La prima cifra significativa di v è uguale alla prima cifra significativa di w (poiché la moltiplicazione per una potenza intera di 10 si limita, eventualmente, a traslare le cifre dello sviluppo decimale di v). Se ad esempio $v = \pi^7 = 3020,29323\ldots$, allora $v = 10^3 w$ e $w = 3,02029323\ldots$ sta tra 3 e 4. Dunque, dire che la prima cifra significativa di v è uguale a $k \in \{1, 2, \ldots, 9\}$ equivale ad affermare che

$$k \leq w < k + 1$$

e quindi

$$\log_{10}(k) \leq \log_{10}(w) < \log_{10}(k+1).$$

Poiché $\log_{10}(v) = M + \log_{10}(w)$ e $0 \leq \log_{10}(w) \leq 1$ abbiamo però

$$\langle \log_{10}(v)\rangle = \log_{10}(w)$$

e quindi

$$\log_{10}(k) \leq \langle \log_{10}(v)\rangle < \log_{10}(k+1)\ .$$

Newcomb ha scritto che ad essere equiprobabili non sono le 9 possibili "prime cifre significative" di un generico numero positivo v, ma le mantisse $\langle \log_{10}(v)\rangle$. Quindi, per qualsiasi intervallo $[a, b)$ contenuto in $[0, 1)$ la probabilità che $\langle \log_{10}(v)\rangle$ appartenga ad $[a, b)$ deve essere uguale alla lunghezza $b - a$ di questo intervallo. In particolare, per la disuguaglianza precedente, la probabilità che la prima cifra significativa di v sia uguale a k deve essere uguale alla lunghezza

$$\log_{10}(k+1) - \log_{10}(k) = \log_{10}(1 + 1/k)$$

PERCENTAGE OF TIMES THE NATURAL NUMBERS 1 TO 9 ARE USED AS FIRST DIGITS IN NUMBERS, AS DETERMINED BY 20,229 OBSERVATIONS

Group	Title	First Digit									Count
		1	2	3	4	5	6	7	8	9	
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	$n^1, \sqrt{n}, \cdots$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	*Digest*	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n!, n^n \cdots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Probable Error		±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3	—

dell'intervallo $[\log_{10}(k), \log_{10}(k+1)]$. Scriviamo i valori numerici delle lunghezze di questi intervalli.

$$\log_{10}(2/1) = 0,30103\ldots \qquad \log_{10}(3/2) = 0,17609\ldots \qquad \log_{10}(4/3) = 0,12494\ldots$$
$$\log_{10}(5/4) = 0,09691\ldots \qquad \log_{10}(6/5) = 0,079181\ldots \qquad \log_{10}(7/6) = 0,066947\ldots$$
$$\log_{10}(8/7) = 0,057992\ldots \qquad \log_{10}(9/8) = 0,051153\ldots \qquad \log_{10}(10/9) = 0,045757\ldots$$

Da questo Newcomb sembra avere concluso che la probabilità che la prima cifra sia 1 è circa il $30,1\%$, la probabilità che la prima cifra sia 2 è circa il $17,6\%$, etc.

Ovviamente non abbiamo scoperto nulla e tantomeno dimostrato la validità di una "legge delle prime cifre". Abbiamo solo verificato che se consideriamo una successione di numeri e supponiamo che le mantisse dei loro logaritmi (in base 10) siano equiprobabili (nel senso visto sopra), allora la percentuale di prime cifre significative uguali a k è circa uguale a $\log_{10}(1 + 1/k)$.

Questo fenomeno fu studiato nuovamente nel 1938 dal fisico e ingegnere elettrico Frank Benford, apparentemente ignaro della nota di Newcomb, che in un articolo sui *Proceedings of the American Philosophical Society* ([1]) presentò numerose successioni numeriche (aree di fiumi, popolazioni, indirizzi, ma anche potenze dei numeri interi, fattoriale, ...) che, soprattutto quando considerate insieme, fornivano una buona evidenza alla "legge logaritmica" descritta sopra, della quale neppure Benford forniva una giustificazione, e che da quel momento fu associata al suo nome. La figura è presa dall'articolo di Benford e riporta i dati da lui raccolti.

Come verifica empirica consideriamo le popolazioni dei comuni italiani. A fronte di 8095 comuni abbiamo (se per ogni $k = 1, 2, \ldots, 9$ indichiamo con $C(k)$ il numero dei comuni il cui numero di abitanti inizia con la cifra k)

$$
\begin{aligned}
&C(1) = 2482 \approx 30,661\%, && \text{confrontiamo con } \log_{10}(2) = 0,30103\ldots\\
&C(2) = 1376 \approx 16,998\%, && \text{confrontiamo con } \log_{10}(3/2) = 0,17609\ldots\\
&C(3) = 1032 \approx 12,749\%, && \text{confrontiamo con } \log_{10}(4/3) = 0,12494\ldots\\
&C(4) = 792 \approx 9,7838\%, && \text{confrontiamo con } \log_{10}(5/4) = 0,09691\ldots\\
&C(5) = 633 \approx 7,8196\%, && \text{confrontiamo con } \log_{10}(6/5) = 0,079181\ldots\\
&C(6) = 533 \approx 6,5843\%, && \text{confrontiamo con } \log_{10}(7/6) = 0,066947\ldots\\
&C(7) = 473 \approx 5,8431\%, && \text{confrontiamo con } \log_{10}(8/7) = 0,057992\ldots\\
&C(8) = 428 \approx 5,2872\%, && \text{confrontiamo con } \log_{10}(9/8) = 0,051153\ldots\\
&C(9) = 346 \approx 4,2742\%, && \text{confrontiamo con } \log_{10}(10/9) = 0,045757\ldots
\end{aligned}
$$

Le popolazioni dei comuni italiani costituiscono un campione ragionevolmente grande e in questo caso le legge di Benford è rispecchiata molto fedelmente. Ora consideriamo un campione più piccolo fatto da 250 elementi: le superfici degli stati della terra.

Questa volta le previsioni della legge di Benford sono rispecchiate meno fedelmente, ma la tendenza alla "legge logaritmica" è comunque evidente.

$$
\begin{aligned}
&C(1) = 71 \approx 28,4\%\\
&C(2) = 48 \approx 19,2\%\\
&C(3) = 30 \approx 12\%\\
&C(4) = 25 \approx 10,4\%\\
&C(5) = 21 \approx 8,4\%\\
&C(6) = 16 \approx 6,4\%\\
&C(7) = 18 \approx 7,2\%\\
&C(8) = 7 \approx 2,8\%\\
&C(9) = 13 \approx 5,2\%
\end{aligned}
$$

Si potrebbe pensare che qualunque insieme di numeri purché sufficientemente numeroso debba soddisfare la legge di Benford. Le cose non stanno però così: utilizzando Excel è abbastanza semplice generare un insieme di numeri casuali e verificare che ognuna della cifre da 1 a 9 compare come prima cifra significativa con frequenza 1/9. Questo significa che a volte i dati numerici del mondo reale sono meno casuali (o meglio, diversamente casuali) rispetto a quello che ci aspettiamo. Guardare solo i "dati del mondo reale" appare però insufficiente se pensiamo che anche alcune successioni numeriche come 2^n, o $n!$, o la successione di Fibonacci soddisfano la condizione di Benford.

La successione di Fibonacci
Questa successione è definita da $F_0 = 1$, $F_1 = 1$, e per ogni intero $n \geq 2$ da $F_n = F_{n-2} + F_{n-1}$. I numeri di Fibonacci hanno la forma esplicita

$$F_n = \frac{1}{\sqrt{5}} \left\{ \left(\frac{1+\sqrt{5}}{2} \right)^n - \left(\frac{1-\sqrt{5}}{2} \right)^n \right\}$$

e sono stati introdotti in occidente da Leonardo Fibonacci all'inizio del 1200 per studiare un problema sulla riproduzione dei conigli.

Definiamo ora meglio cosa intendiamo per condizione di Benford.

Definizione 1. *Diciamo che una successione reale positiva $\{t_n\}$ è una successione di Benford (in base 10) se*

$$\lim_{N \to +\infty} \frac{\operatorname{card}\{n \leq N : \text{la prima cifra non nulla di } t_n \text{ è uguale a } k\}}{N} = \log_{10}\left(1 + \frac{1}{k}\right) ,$$

dove $\operatorname{card} A$ *indica il numero degli elementi contenuti in un insieme finito* A.

Questa definizione significa che, per N molto grande la percentuale dei numeri t_n con $n \leq N$ e la prima cifra uguale ad 1 si avvicina al $30,103\ldots\%$, quella dei numeri t_n con la prima cifra uguale a 2 si avvicina al $17,609\ldots\%$, etc.

2 Successioni uniformemente distribuite

Per andare avanti dobbiamo presentare la definizione di successione uniformemente distribuita, introdotta da Hermann Weyl nel 1916 e legata alla legge forte dei grandi numeri (vedi [5], [12, 7.4]).

La legge forte dei grandi numeri
Mettiamo in un'urna 10 palline con i numeri $0, 1, 2, \ldots, 9$. Ripetiamo infinite volte l'operazione di estrarre una pallina, segnarne il valore e poi rimetterla nell'urna. Otteniamo così una successione infinita

$$\{\omega_1, \omega_2, \omega_3, \ldots\}$$

dove ciascun ω_j assume con probabilità $1/10$ ciascuno dei valori $0, 1, 2, \ldots, 9$. A questa successione associamo il numero $\omega = 0, \omega_1\omega_2\omega_3\ldots$ che è un numero reale compreso tra 0 e 1. Se tralasciamo le successioni in cui ω_n è definitivamente uguale a 9 otteniamo una corrispondenza biunivoca in cui ad una successione infinita di estrazioni corrisponde uno ed un solo numero reale compreso tra 0 e 1. In questo modo la misura sull'intervallo $[0, 1]$ diventa una misura sull'insieme delle successioni di estrazioni. La Legge forte dei grandi numeri (dimostrata tra il 1909 e il 1916 da Émile Borel e Francesco Cantelli) implica che, se N è molto grande, quasi certamente (nel senso della misura) ciascuna cifra tra 0 e 9 apparirà circa $1/10$ delle volte. Questo significa che prendendo un numero a caso nell'intervallo $[0, 1]$ questo numero, con probabilità 1, ha (nel senso del limite per $N \to +\infty$) uguale porzione (cioè $1/10$) di cifre $0, 1, 2, \ldots, 9$. Si può passare da una singola cifra ad una qualsiasi sequenza finita di cifre: un numero a caso nell'intervallo $[0, 1]$ contiene con probabilità 1 il numero di cellulare del lettore infinite volte e con la frequenza dovuta (o, se vogliamo, con probabilità 1 una scimmia che batta a caso sui tasti scriverà la Divina Commedia infinite volte e, nel senso del limite, con la frequenza dovuta). Si può dimostrare che questa proprietà equivale al fatto che spostando indietro di n passi la virgola di quasi ogni numero reale tra 0 e 1, la successione delle parti frazionarie via via ottenute è uniformemente distribuita.

Definizione 2. *Una successione* $\{t_n\}$ *a valori nell'intervallo* $[0, 1)$ *è uniformemente distribuita se per ogni* $0 \leq a < b < 1$ *si ha*

$$\lim_{N\to+\infty} \frac{\operatorname{card}\{n \leq N : a \leq t_n < b\}}{N} = b - a\ .$$

Cioè una successione di numeri in $[0, 1)$ è uniformemente distribuita se per N molto grande la percentuale dei numeri t_n che cadono in un generico intervallo $[a, b)$ si avvicina alla lunghezza $b - a$ di questo intervallo. Ad esempio, la *successione di van der Corput*

$$\{t(j)\}_{j=1}^{\infty} = \left\{\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \frac{7}{8}, \frac{1}{16}, \frac{9}{16}, \frac{5}{16}, \frac{13}{16}, \frac{3}{16}, \frac{11}{16}, \frac{7}{16}, \frac{15}{16}, \ldots\right\}\ ,$$

è uniformemente distribuita (per ogni intero positivo $j = \sum a_k 2^k$, scritto in base 2, si definisce $t(j) = \sum a_k 2^{-k-1}$).

Una successione uniformemente distribuita è utile per "campionare" una funzione su $[0,1]$ della quale dobbiamo stimare l'integrale. Infatti un modo molto ragionevole di definire (anche agli studenti, vedi [4]) l'integrale di Riemann di una opportuna funzione f definita sull'intervallo $[0,1]$ è

$$\int_0^1 f(x)\ dx \overset{\text{definizione}}{=} \lim_{N\to+\infty} \left(\frac{1}{N} \sum_{j=1}^{N} f(t_j) \right) ,$$

dove per ogni intero j tra 0 e $N-1$ il punto t_j è scelto liberamente nell'intervallo $\left[\frac{j}{N}, \frac{j+1}{N}\right)$ e il limite non deve dipendere da questa scelta.

In questo modo lo studente vede confrontate le due nozioni di funzione più familiari. Da un lato quella di *tabella di valori:*

$$\begin{array}{cc} t_1 & f(t_1) \\ t_2 & f(t_2) \\ t_3 & f(t_3) \\ t_4 & f(t_4) \\ \vdots & \vdots \\ t_N & f(t_N) \end{array}$$

con la facile e utile operazione di *media aritmetica:*

$$\frac{1}{N} \sum_{j=1}^{N} f(t_j) = \frac{f(t_1) + f(t_2) + \ldots + f(t_N)}{N} .$$

Dall'altro quella di funzione definita su un intervallo reale, con la meno facile operazione di integrazione

$$\int_0^1 f(x)\ dx .$$

La somma $\frac{1}{N} \sum_{j=1}^{N} f(t_j)$ è detta *somma di Riemann* e - particolarmente nei problemi con molte variabili, non rari in fisica e in finanza matematica - fornisce, per opportune scelte dei t_j, buone stime di integrali altrimenti difficilmente trattabili.

La Definizione 1 e la Definizione 2 sembrano, ed in effetti sono, molto vicine. Si può dimostrare che se $\{t_n\}$ è una successione positiva infinita e la successione $\{\langle \log_{10}(t_n) \rangle\}$ delle parti frazionarie di $\log_{10}(t_n)$ è uniformemente distribuita, allora $\{t_n\}$ è di Benford (è quello che abbiamo fatto quando abbiamo discusso il brano di Newcomb). Più in generale si può definire una successione di Benford **forte** chiedendo non solo che la prima cifra soddisfi al legge di Benford, ma che ogni sequenza finita $u_1 u_2 \ldots u_r$ di cifre decimali (che non inizi con 0) appaia con la frequenza dovuta, che è

$$\log_{10}(u_1 u_2 \ldots u_r + 1) - \log_{10}(u_1 u_2 \ldots u_r) = \log_{10}\left(1 + \frac{1}{u_1 u_2 \ldots u_r}\right) .$$

Si dimostra allora che $\{t_n\}$ è una successione forte di Benford se e solo se $\{\langle \log_{10}(t_n) \rangle\}$ è uniformemente distribuita.

Un celebre teorema di Lepold Kronecker dice che se α è un numero irrazionale, allora la successione $\{\langle \alpha n \rangle\}$ delle parti frazionarie della progressione aritmetica αn è uniformemente distribuita sull'intervallo $[0, 1)$. Ad esempio, se $\alpha = \sqrt{2}$, allora la successione

$$\begin{aligned}
\left\langle \sqrt{2} \right\rangle &= 0,41414\ldots \\
\left\langle 2\sqrt{2} \right\rangle &= 0,82842\ldots \\
\left\langle 3\sqrt{2} \right\rangle &= 0,24264\ldots \\
\left\langle 4\sqrt{2} \right\rangle &= 0,65685\ldots \\
\left\langle 5\sqrt{2} \right\rangle &= 0,07106\ldots \\
\left\langle 6\sqrt{2} \right\rangle &= 0,48528\ldots \\
\left\langle 7\sqrt{2} \right\rangle &= 0,89949\ldots \\
\left\langle 8\sqrt{2} \right\rangle &= 0,31371\ldots \\
&\vdots
\end{aligned}$$

è uniformemente distribuita nell'intervallo $[0, 1)$.

Da questo si deduce che la successione $\{2^n\}$ delle potenze di 2 soddisfa la legge (forte) di Benford. Basta infatti dimostrare che la successione $\{\langle \log_{10}(2^n) \rangle\} = \{\langle n \log_{10}(2) \rangle\}$ è uniformemente distribuita, ma questo segue dal teorema di Kronecker, poiché $\log_{10}(2)$ è irrazionale (se fosse $\log_{10} 2 = p/q$ avremmo $2 = 10^{p/q}$, cioè $2^q = 10^p$, cioè $2^{q-p} = 5^p$, che è impossibile). Come verifica senza pretese (poiché non abbiamo detto nulla sulla *velocità* con cui le prime cifre vanno a soddisfare la legge di Benford) scriviamo le prime cento potenze di 2,

$$\begin{aligned}
2^1 &= 2 \\
2^2 &= 4 \\
2^3 &= 8 \\
&\vdots \\
2^{99} &= 633825300114114700748351602688 \\
2^{100} &= 1267650600228229401496703205376
\end{aligned}$$

e osserviamo che tra di esse

30 iniziano con la cifra 1,
17 iniziano con la cifra 2,
13 iniziano con la cifra 3,
10 iniziano con la cifra 4,
7 iniziano con la cifra 5,
7 iniziano con la cifra 6,
6 iniziano con la cifra 7,
5 iniziano con la cifra 8,
5 iniziano con la cifra 9.

In modo abbastanza simile si dimostra che anche la successione di Fibonacci $\{F_n\}$ è di Benford. Provare che $n!$ è di Benford non è molto diverso, ma richiede la formula di Stirling ($n! \sim \sqrt{2\pi n} n^n e^{-n}$ per $n \to +\infty$).

Si può dimostrare (vedi [12, p.122]) che la successione $\{\langle \log n \rangle\}$ non è uniformemente distribuita in $[0,1)$. Da questo segue che $\{n\}$ non è una successione forte di Benford, ma non possiamo dedurre che non è una successione di Benford (cioè che non soddisfa la legge della *sola* prima cifra). Mostriamo direttamente che $\{n\}$ non è una successione di Benford calcolando, per ogni intero positivo N quanti sono gli interi positivi minori od uguali ad N che iniziano con la cifra 1 e dividendo questo numero per N. Chiamiamo $q(N)$ il risultato.

$$\begin{array}{ll}
\text{se } N = 1 & \text{allora } q(N) = 1 \\
\text{se } 1 < N < 10 & \text{allora } q(N) = 1/N \\
\text{se } 10 \leq N < 20 & \text{allora } q(N) = 1 - 8/N \\
\text{se } 20 \leq N < 99 & \text{allora } q(N) = 11/N \\
\text{se } 100 \leq N < 200 & \text{allora } q(N) = 1 - 88/N \\
\text{se } 200 \leq N < 1000 & \text{allora } q(N) = 111/N \\
\vdots & \vdots \\
\text{se } 10^k \leq N < 2 \cdot 10^k & \text{allora } q(N) = 1 - \frac{8}{9}\frac{10^k - 1}{N} \\
\text{se } 2 \cdot 10^k \leq N < 10^{k+1} & \text{allora } q(N) = \frac{10^{k+1} - 1}{9N}
\end{array}$$

Dunque, per $N = 2 \cdot 10^k - 1$ abbiamo

$$c(N) = 1 - \frac{8}{9}\frac{(N-1)}{2N} \longrightarrow \frac{5}{9} ,$$

mentre per $N = 10^{k+1} - 1$ abbiamo

$$c(N) = \frac{1}{9} .$$

Si verifica allora che

$$\liminf_{N \to +\infty} c(N) = \frac{1}{9} , \quad \limsup_{N \to +\infty} c(N) = \frac{5}{9} ,$$

e quindi $\{n\}$ non è una successione di Benford.

Ci aspettiamo che la legge di Benford non dipenda dall'unità di misura usata per stimare aree o altre grandezze fisiche. Tornando ai dati sulla superficie degli stati della terra, questo significa, ad esempio, che la distribuzione delle prime cifre non deve cambiare troppo se come unità di misura si utilizzando le miglia quadrate invece che i chilometri quadrati. Indichiamo con S la superficie di un dato stato misurata in chilometri quadrati e con S^* la superficie dello stesso stato espressa in miglia quadrate. Supponiamo che la prima cifra di S sia 1. Allora la prima cifra di S^* è 2 oppure 3 (e viceversa, se la prima cifra di S^* è 2 oppure 3, allora la prima cifra di S è 1). Infatti

$$\log_{10}(2) - \log_{10}(1) = \log_{10}(4) - \log_{10}(2) ,$$

e analogamente per le altre cifre. Più in generale, si può dimostrare l'invarianza delle successioni di Benford rispetto ai cambi di scala (vedi [11], cioè se $\{t_n\}$ soddisfa la legge di Benford forte, allora, per ogni numero reale $\alpha > 0$, anche $\{\alpha t_n\}$ soddisfa la legge di Benford forte. Si può dimostrare anche il viceversa: le successioni che (con un'opportuna definizione) sono invarianti per cambi di scala soddisfano la legge di Benford forte (vedi [3]).

3 Applicazioni

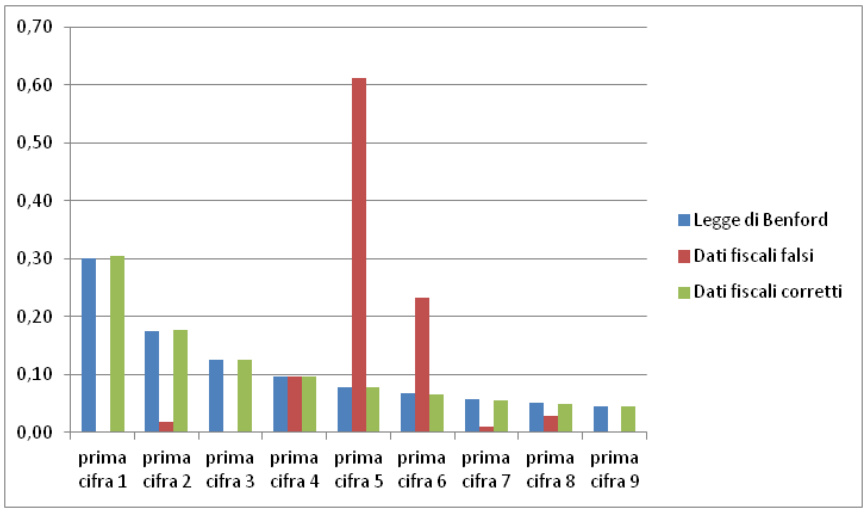
La legge di Benford ha alcune applicazioni semplici e nello stesso tempo molto efficaci. Per introdurne qualcuna torniamo alle considerazioni fatte sui comuni italiani. Se le loro popolazioni seguono molto bene la legge di Benford, possiamo assumere che la seguiranno anche

in futuro. Abbiamo così una tecnica per valutare l'attendibilità di modelli demografici: basta stabilire se i dati previsti (pur approssimati) seguono la legge di Benford. Più in generale, la legge di Benford riceve attenzione da chi deve investigare sulle frodi che coinvolgono una considerevole quantità di dati numerici, come ad esempio le frodi fiscali (il US Internal Revenue Service usa la legge di Benford per evidenziare le dichiarazioni dei redditi sospette), assicurative o relative alle richieste di rimborso presentate a compagnie pubbliche o private dai propri dipendenti (vedi [8], [10], [9]). Per studiare questi dati si può esaminare la prima o le prime cifre dei numeri dichiarati (considerando cioè situazioni intermedie tra la legge di Benford e la legge di Benford forte). L'applicazione della legge di Benford nella ricerca delle frodi può sembrare una "pistola con un solo colpo", inutile nel momento in cui la legge diventa conosciuta al grosso pubblico. In realtà la sua applicazione può variare in modalità e raffinatezza, rendendo comunque difficile la creazione di dati numerici falsi, come sostiene in questo commento Mark Nigrini (vedi [6]).

> Il problema di quelli che commettono frodi è che fino al momento in cui tutti i dati sono inseriti non hanno idea di come appare il quadro complessivo. Le frodi di solito riguardano una parte di un dataset, ma quelli che frodano non sanno come questo insieme sarà analizzato: per trimestre, per dipartimento, o per regione. Verificare che la frode non viola la legge di Benford diventa duro - e molti di quelli che frodano non sono ingegneri aerospaziali.

La tabella che segue, dovuta a Mark Nigrini, confronta le percentuali di prime cifre uguali a $k \in \{1, 2, \ldots, 9\}$ secondo

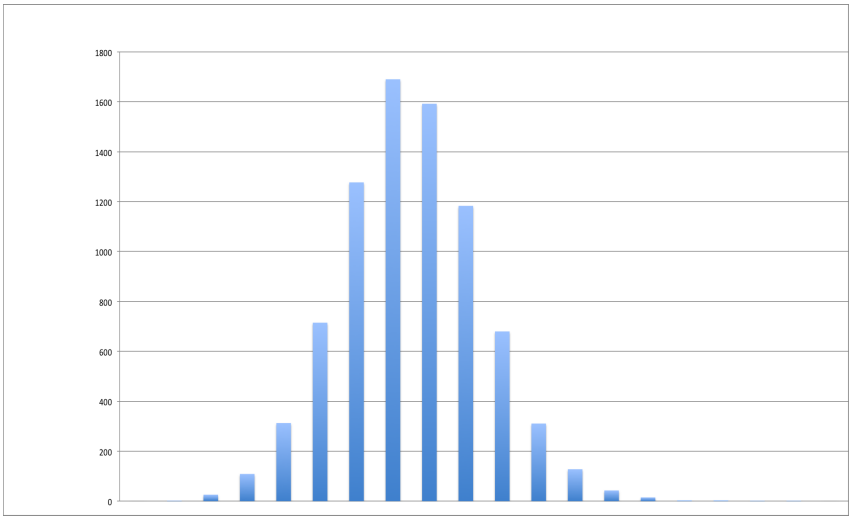
1. la legge di Benford
2. dati fiscali falsi dell'anno 1995 nella Contea di Kings, New York
3. dati fiscali corretti



Altre applicazioni della legge di Benford comprendono la diagnosi di modelli matematici in biologia e in finanza, la scoperta di immagini digitali alterate, la suddivisione di una memoria per l'allocazione di dati.

4 Giustificazioni

Chi intende usare la legge di Benford vorrebbe averne una spiegazione convincente per sapere quando è applicabile a dati numerici reali. A questa domanda non è ancora stata data una riposta completamente soddisfacente (si veda ad esempio [2]). Il motivo per cui $\{2^n\}$, $\{n!\}$ o la successione di Fibonacci soddisfano la legge di Benford è forse diverso dal motivo per cui la soddisfano il numero di abitanti dei comuni italiani o la superficie dei paesi nel mondo. È abbastanza ovvio che i dati devono essere molti e distribuiti su più ordini di grandezza; le altezze delle persone ad esempio non possono soddisfare la legge di Benford perché iniziano quasi tutte con la cifra 1. Sia il caso delle popolazioni dei comuni italiani sia quello delle superfici dei paesi del mondo forniscono invece numeri che spaziano su più ordini di grandezza. Nel primo caso si va infatti dalle poche decine di abitanti dei piccoli comuni ai milioni delle città metropolitane mentre nel secondo caso si va dai pochi chilometri quadrati della Città del Vaticano alle decine di milioni di chilometri quadrati della Russia. Supponiamo ora di voler rappresentare con un grafico le frequenze con cui si presentano le popolazioni dei comuni. Occorre quindi dividere il numero di abitanti in classi e per ogni classe conteggiare quanti comuni hanno un numero di abitanti che ricade nella classe considerata. Da subito ci si rende però conto che è impossibile ottenere dei grafici significativi utilizzando classi equispaziate. Se ad esempio utilizziamo una ampiezza della classe pari a 5000, vi sono 5683 comuni con un numero di abitanti compreso tra 0 e 5000, 1192 con un umero di abitanti compreso tra 5001 e 10000 e 480 con un numero di abitanti compreso tra 10001 e 15000. Le prime tre classi contengono pertanto 7355 comuni su 8092. Per arrivare però alla popolazione di Roma (2.761.477 abitanti) abbiamo bisogno di 554 classi (la maggior parte delle quali vuote). La natura dei dati, ed in particolare il fatto che spaziano su più ordini di grandezza, ci costringe pertanto ad utilizzare classi non equispaziate. Utilizzando delle classi di ampiezza progressivamente crescente, come ad esempio $10-20$, $20-40$, $40-80$, $80-160$, ecc. otteniamo il seguente grafico:



In questa scala "logaritmica" (nel senso che non sono le classi ad essere equispaziate ma i loro logaritmi) le frequenze delle popolazioni dei comuni hanno una distribuzione normale; si concentrano cioè attorno al valore più frequente (che nel nostro caso corrisponde ai comuni con circa 2500 abitanti) per poi diminuire velocemente quando ci si allontana da questo valore con il tipico andamento a campana. Si può dimostrare che più la campana è allargata, meglio i dati soddisfano la legge di Benford.

Riferimenti bibliografici

[1] F. Benford, *The Law of Anomalous Numbers,* Proc. Am. Philos. Soc., **78** (1938), 551-572.

[2] A. Berger e T. Hill, *Benford's Law Strikes Back: No Simple Explanation in Sight for Mathematical Gem*, The Mathematical Intelligencer, 33 (2011), 85-91.

[3] A. Berger e T. Hill, *A basic theory of Benford's law*, Probability Survey, **8** (2011), 1-126.

[4] M. Bramanti, *Una proposta didattica, come e perché insegnare gli integrali*, Emmeciquadro **36** (2009) 47–68.

[5] L. Kuipers e H. Niederreiter, *Uniform distribuition of sequences,* Dover, 2006.

[6] R. Matthews, *The power of one*, New Scientist, 10 July 1999.

[7] S. Newcomb, *Note on the Frequency of Use of the Different Digits in Natural Numbers*, Am. J. Math., **4** (1881), 39-40.

[8] M. Nigrini, *I've Got Your Number,* J. Accountancy, www.journalofaccountancy.com/Issues/1999/May/nigrini

[9] M. Nigrini, Benford's law, John Wiley & Sons, 2012.

[10] M. Nigrini e L. Mittermaier, *The use of Benford's Law as an Aid in Analytical procedures,* Auditing - A Journal of Practice & Theory **16** (1997), 52-67.

[11] R. Pinkham, *On the distribution of first significant digits,* Ann. Math. Stat., **32** (1961), 1223-1230.

[12] G. Travaglini, *Appunti su teoria dei numeri, Analisi di Fourier e distribuzione di punti,* Unione Matematica Italiana - Pitagora, 2010.

Luca Brandolini, Dipartimento di Ingegneria dell'Informazione e Metodi Matematici, Università di Bergamo, Viale Marconi 5, 24044 Dalmine, Bergamo.
LUCA.BRANDOLINI@UNIBG.IT

Giancarlo Travaglini, Dipartimento di Statistica, Edificio U7, Università di Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano.
GIANCARLO.TRAVAGLINI@UNIMIB.IT