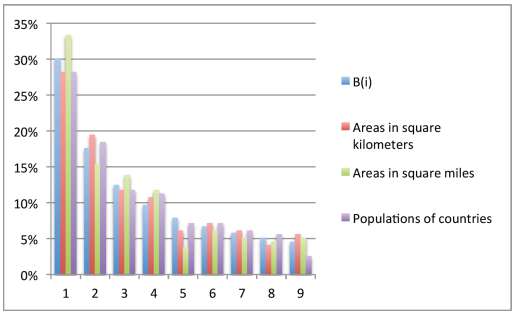
La legge di Benford: imparare a intercettare le frodi

Christiane Rousseau

22 Giugno 2012



È molto rischioso cambiare troppi numeri in certe affermazioni finanziarie se non si conosce un po' di matematica. Infatti, molto spesso i numeri che compaiono in affermazioni finanziarie seguono una certa strana regola matematica, chiamata legge di Benford o legge della prima cifra significativa. Se non si segue la regola allora i numeri falliranno certi test statistici e sarà probabile che siano analizzati con cura.

La legge di Benford afferma che se si collezionano numeri in maniera casuale e si calcolano le frequenze delle loro prime cifre significative, i numeri con 1 come prima cifra significativa dovrebbero apparire circa il 30% delle volte, mentre i numeri con 9 come prima cifra significativa appaiono solo 4.5% delle volte. Questa regola si può osservare in molte altre serie di numeri come le potenze di 2 e i numeri di Fibonacci.

Perché?

Ora, abbiamo spiegazioni soddisfacenti che stiamo per condividere con voi.

La legge di Benford riguarda la distribuzione delle prime cifre significative dei numeri. La prima cifra significativa di un numero positivo è la cifra non nulla più a sinistra della sua espressione decimale. Per esempio, la prima cifra significativa di π è 3, quella di 2371.5 è 2 e quella di 0.00563 è 5. Un altro modo per definirla, che sarà utile per la nostra discussione matematica, è

scrivere un numero reale positivo x come un numero $m \in [1, 9)$ moltiplicato per una potenza di 10:

$$x = m10^n, \quad n \in \mathbb{Z}$$

Quindi la prima cifra significativa di x è la parte intera di m, che può essere denotata con $\lfloor m \rfloor$. Il numero m è chiamata la *mantissa* di x. Affermiamo, ora, che se si collezionano numeri in modo casuale e si calcola la frequenza $B(i)$ della prima cifra significativa i, allora $B(i)$ è data approssimativamente da $log_{10}(1 + \frac{1}{i})$. Questo fornisce le frequenze:

i	1	2	3	4	5	6	7	8	9
$B(i)$	0.3010	0.1761	0.1249	0.0969	0.0792	0.0669	0.0580	0.0511	0.0458

Figura 1: Frequenze della legge di Benford

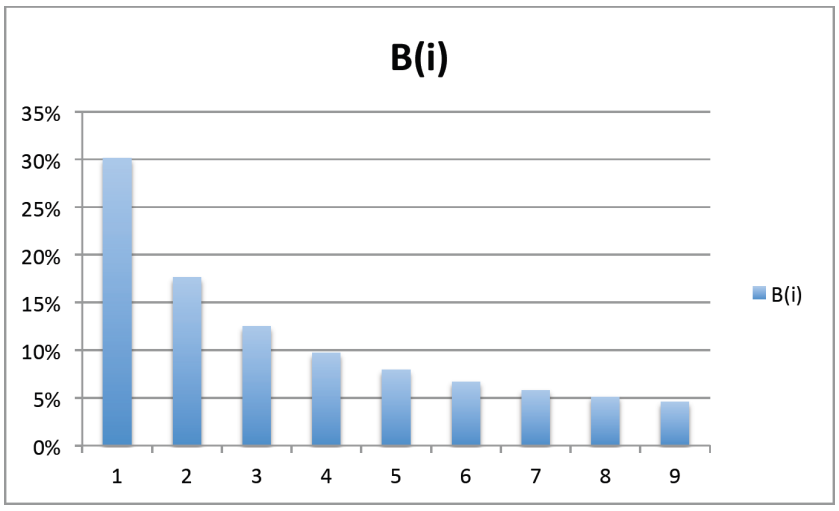


Figura 2: Frequenze B(i) della legge di Benford

Diamo ora una breve nota storica. Il fenomeno fu scoperto per la prima volta dall'astronomo Simon Newcombe (1835-1909) che notò che le prime pagine delle tavole logaritmiche corrispondenti alle prime cifre significative piccole sembravano molto più consumate delle ultime pagine. La sua scoperta fu dimenticata e la legge fu riscoperta da Frank Benford (1883-1948) intorno al 1938. Frank Benford collezionò decine di migliaia di numeri da tutte le origini che seguono la sua legge. Il data base moderno di Simon Plouffe che contiene 215 milioni di costanti matematiche segue anch'esso la legge di Benford.

Anche molti insiemi di numeri che non sono casuali seguono la legge di Benford. Questo è il caso delle popolazioni di nazioni, delle aree di nazioni, delle

lunghezza dei fiumi ... Forse potreste fermarmi e iniziare ad essere scettici... In quali unità di misura sono collezionate queste lunghezze ed aree? Le lunghezze sono espresse in miglia o in chilometri? **Questo non ha importanza...** Se le lunghezze dei fiumi in chilometri seguono la legge di Benford, allora le lunghezze in miglia seguono la legge di Benford! Un cambiamento di unità corrisponde a un cambiamento di scala. Vedremo che la legge di Benford è **invariante per cambiamenti di scala**. Inoltre, è l'unica legge di probabilità invariante per cambiamenti di scala.

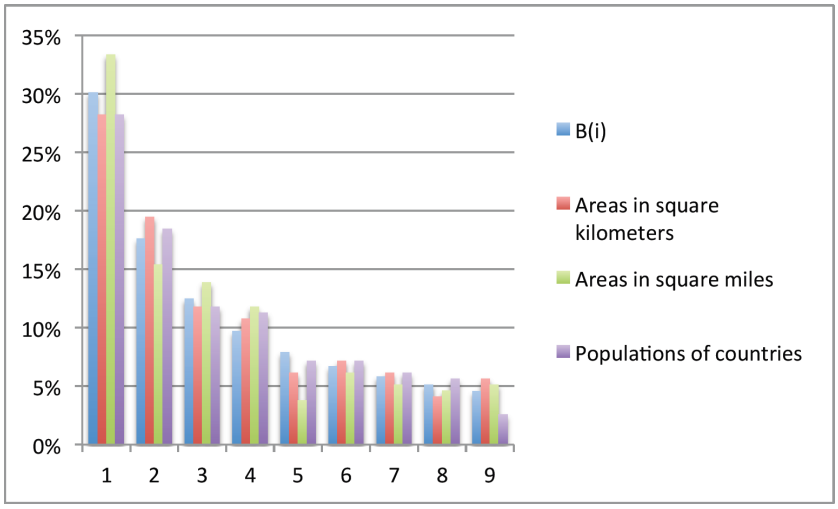


Figura 3: Alcuni dati che approssimativamente seguono la legge di Benford: le aree delle nazioni in chilometri quadrati, le aree delle nazioni in miglia al quadrato e le popolazioni delle nazioni

Nell'introduzione, vi ho detto che i numeri di Fibonacci seguono la legge di Benford. Ma, in un certo senso, la legge di Benford è soggettiva, poiché dipende dalla base 10 in cui scriviamo i numeri. In una qualche base b con $b \neq 10$, le cifre non nulle sono degli elementi dell'insieme $\{1; ....; b-1\}$ e la legge di Benford in base b afferma che la frequenza della prima cifra significativa i è $B_b(i) = log_b(1 + \frac{1}{i})$. Bene! I numeri di Fibonacci seguono la legge di Benford in una qualunque base b! La legge di Benford è **invariante per cambiamenti di base**. È la sola legge di probabilità non banale invariante per cambiamenti di base.

È tempo di dare spiegazioni. Esse richiederanno che vi ricordiate qualcosa del vostro corso di probabilità . Ma potreste preferire di sperimentare inizialmente da voi stessi prima di iniziare a leggere della matematica più seria.

1. Invarianza sotto il cambiamento di scala

Consideriamo un semplice cambiamento di scala ottenuto moltiplicando tutti i numeri di un insieme di numeri per 2. Se consideriamo i numeri con 1 come prima cifra significativa, allora essi vengono cambiati in numeri con 2 o 3 come prima cifra significativa. È facile verificare che $B(1) = B(2) + B(3)$. Infatti,

$$\begin{aligned} B(2) + B(3) &= log_{10}(1 + \frac{1}{2}) + log_{10}(1 + \frac{1}{3}) \\ &= log_{10}\frac{3}{2} + log_{10}\frac{4}{3} = log_{10}\frac{3}{2}\frac{4}{3} \\ &= log_{10}2 = B(1) \end{aligned}$$

In modo analogo, potreste verificare che $B(2) = B(4) + B(5)$, etc. Ma, come ci si comporta se si passa dalle miglia ai chilometri, cioè se si moltiplicano i numeri per 1.6? Come affermato, la legge di Benford è troppo restrittiva e occorre generalizzarla. Che cosa significa che la prima cifra significativa è i? Significa che la mantissa m appartiene all'intervallo $[i, i + 1)$. Quindi, la legge di Benford è una distribuzione parziale di probabilità della mantissa. La legge di Benford generalizzata alla mantissa (che, con abuso di linguaggio, chiameremo legge di Benford) è data da una funzione di densità sull'intervallo $[1, 10)$.
Quando prendiamo un numero casuale, possiamo calcolarne la sua mantissa. Questo ci dà una variabile casuale M che assume valori in $[1, 10)$. Diciamo che essa segue la legge di Benford se la sua funzione di densità è data da

$$f(x) = \begin{cases} \frac{1}{xlog_{10}}, & x \in [1, 10) \\ 0, & altrimenti \end{cases}$$

Se $P(a \leq M < b)$ rappresenta la probabilità che $a \leq M < b$, allora ciò significa che dobbiamo avere

$$P(a \leq M < b) = \int_a^b f(x)dx$$

È davvero una generalizzazione della legge di Benford, poiché

$$\begin{aligned} B(i) = P(i \leq M < i + 1) &= \int_i^{i+1} \frac{1}{xlog_{10}}dx \\ &= \frac{1}{log_{10}}(log(i + 1) - log(i)) = \frac{1}{log_{10}}(log\frac{i+1}{i}) \\ &= \frac{log(1 + \frac{1}{i})}{log10} = log_{10}(1 + \frac{1}{i}) \end{aligned}$$

Che cosa vuol dire che una variabile casuale X su $[1, 10)$ è invariante per cambiamenti di scala? Significa che, se c è un numero reale positivo e se consideriamo la variabile casuale $Y = cX$, allora la mantissa M della variabile casuale Y ha la stessa funzione di densità di X. Non è difficile mostrare che questo è il caso di quando X segue la legge di Benford, ma ci sono molti casi da distinguere che dipendono dalla grandezza di c. Vedremo uno di questi casi e vi lasceremo da fare gli altri. Possiamo scrivere $c = m10^r$, con $m \in [1, 10)$ la mantissa di c. Poiché la mantissa di cX è la stessa di mX, è sufficiente considerare il caso $c \in [1, 10)$.
Qual è lo strumento per mostrare ciò? Potreste ricordare dal vostro corso di probabilità che, per una variabile casuale continua, la funzione di distribuzione (cumulativa) è talvolta più utile della funzione di densità per una variabile casuale continua. La funzione di distribuzione di una variabile casuale M è definita da

$$F(x) = P(M \leq x)$$

Se X segue la legge di Benford, allora la sua funzione di distribuzione è data da

$$F(x) = \begin{cases} 0, & x < 1, \\ log_{10} x, & x \in [1, 10), \\ 1, & x \geq 10. \end{cases} \tag{1}$$

Quindi, dobbiamo mostrare che se X segue la legge di Benford e M è la mantissa di cX per $c \in [1, 10)$, allora la funzione di distribuzione di M è data da (1).

A questo scopo, dobbiamo calcolare $P(M \leq z)$ per $z \in [1, 10]$. M è la mantissa di cX che assume valori nell'intervallo $[c, 10c)$. Quindi $M = cX$, quando $cX < 10$ e $M = cX/10$ quando $cX \geq 10$. Il primo caso accade quando $z < c$. L'unica possibilità affinché la mantissa di cX appartenga all'intervallo $[1, c)$ è che $cX \in [10, 10c]$. Allora la mantissa di cX è uguale a $cX/10$.

Quindi,

$$\begin{aligned} P(M \leq z) =& P(1 \leq cX/10 \leq z) \\ &= P(\frac{10}{c} \leq X \leq \frac{10z}{c}) \\ &= F(\frac{10z}{c}) - F(\frac{10}{c}) \\ &= log_{10}\frac{10z}{c} - log_{10}\frac{10}{c} \\ &= log_{10}z + log_{10}\frac{10}{c} - log_{10}\frac{10}{c} \\ &= log_{10}z, \end{aligned}$$

come atteso. Gli altri casi si trattano allo stesso modo.

Il reciproco è più interessante...

2. La legge di Benford è la sola legge di probabilità sulla mantissa invariante per cambiamenti di scala

Sembra un'affermazione sorprendente! Tuttavia, si vedrà che la sua dimostrazione non è molto più complicata della precedente argomentazione. Sia X una variabile casuale che rappresenta la mantissa e che assume valori nell'intervallo $[1, 10)$. Cerchiamo la sua funzione di distribuzione $F(x)$, sotto l'ipotesi che X sia invariante per cambiamenti di scala. Allora, dobbiamo calcolare

$$F(x) = P(X \leq x) = P(1 \leq X \leq x)$$

Quindi, dobbiamo avere $F(0) = 0$ e $F(10) = 1$.
La difficoltà principale della dimostrazione è capire cosa significhi che X è invariante per cambiamenti di scala. Poiché $1 \leq X \leq x$ e $c \leq cX \leq cx$ sono gli stessi eventi, allora abbiamo che

$$P(1 \leq X \leq x) = P(c \leq cX \leq cx) = F(x). \tag{2}$$

Come prima, consideriamo il caso di un qualche $c \in [1, 10)$ così che $cx < 10$ (c dipende da x). Allora, per $c \leq cX \leq cx$, cX è uguale alla sua mantissa. Poiché X è invariante per cambiamenti di scala, allora la mantissa di cX ha la stessa funzione di distribuzione di X. Quindi,

$$P(c \leq cX \leq cx) = F(cx) - F(c).$$

Combinando quest'ultima relazione con la (2), si nota che $F(x)$ soddisfa a

$$F(x) = F(cx) - F(c), \qquad F(1) = 0, \quad F(10) = 1 \tag{3}$$

a condizione che $c \in [1, 10)$ non sia troppo grande. Dobbiamo trovare F dall' *equazione funzionale* (3). Vediamo come farlo. Se poniamo $c = 1 + \epsilon$ si ha che

$$F(x) = F(x(1 + \epsilon)) - F(1 + \epsilon)$$

che può essere scritto

$$\frac{F(x(1 + \epsilon)) - F(x)}{x\epsilon} = \frac{F(1 + \epsilon) - F(1)}{x\epsilon},$$

poiché $F(1) = 0$. Consideriamo il limite quando $\epsilon \to 0$. Dobbiamo riconoscere in ciascuno dei due membri un quoziente il cui limite è una derivata. In particolare, a primo membro il $\lim_{\epsilon \to 0} \frac{F(x(1 + \epsilon)) - F(x)}{x\epsilon}$ è $F'(x)$, mentre nel membro di destra il $\lim_{\epsilon \to 0} \frac{F(1 + \epsilon) - F(1)}{\epsilon}$ è $F'(1)$. Quindi, abbiamo l'equazione differenziale a variabili separabili

$$F'(x) = \frac{F'(1)}{x}$$

la cui soluzione è $F(x) = F'(1)lnx + C$. Poiché $F(1) = 0$ abbiamo $C = 0$ e poiché $F(10) = 1$, allora $F'(1) = \frac{1}{ln10}$. Quindi, $F(x) = \frac{lnx}{ln10} = log_{10}x$ e abbiamo finito!

3. Perché i numeri collezionati da tutte le origini seguono la legge di Benford?

Una risposta fu fornita da Theodore Hill nel 1995; discuteremo brevemente la sua idea. Sicuramente, non tutti gli insiemi di numeri seguono la legge di Benford. Per esempio, se si considera l'altezza in metri degli esseri umani allora, salvo poche eccezioni, compaiono soltanto le prime cifre significative 1 e 2 e se si converte l'altezza in piedi (un piede corrisponde circa a 30 cm) cambierà la legge di distribuzione della prima cifra significativa. Di conseguenza, questo insieme di numeri non è invariante per cambiamenti di scala. Ma, supponiamo di considerare un insieme grande di numeri presi tra tutte le origini e operiamo un cambiamento di scala. Ci sono diversi sottoinsiemi di numeri con la loro particolare scala. Poiché l'insieme è grande e i numeri provengono da tutte le origini, allora è probabile che siano presenti tutte le differenti scale. Moltiplicare tutti i numeri presenti nell'insieme per una costante positiva induce una permutazione delle scale presenti nel nuovo insieme. Così, in generale, potremmo aspettarci che l'insieme di numeri si comporti come se non avesse alcuna scala speciale. Quindi, esso seguirà la legge di Benford.

Questa è una buona spiegazione per gli insiemi di numeri presi da tutte le origini. Ma non spiega perché le aree delle nazioni, le popolazioni di nazioni o le lunghezze dei fiumi dovrebbero seguire la legge di Benford. Per questo caso discuteremo spiegazioni molto recenti (del 2008!) date da Gauvrit, Delahaye e Fewster. La loro spiegazione è anche valida per grandi insiemi di numeri presi tra tutte le origini.

4. È probabile che insiemi di numeri che comprendono diversi ordini di grandezza seguano la legge di Benford!

Stiamo lavorando in base 10 e abbiamo visto che i numeri positivi x possono essere scritti come $x = m10^n$, dove $m \in [1, 10)$ e $n \in \mathbb{Z}$. Potremmo considerare n come l' ordine di grandezza e dire che esistono diversi ordini di grandezza se esistono diversi valori di n per il nostro insieme di numeri (notiamo che una tale proprietà è invariante per cambiamenti di scala!). Per semplificare la spiegazione, supponiamo che i numeri appartengano all'intervallo $[1, 10^6)$. Quindi, i numeri con 1 come prima cifra significativa sono quelli nell'insieme

$$S_1 = [1, 2) \cup [10, 20) \cup [100, 200) \cup [1000, 2000) \cup \left[10^4, 2 \cdot 10^4\right) \cup \left[10^5, 2 \cdot 10^5\right),$$

e negli insiemi simili S_i con le altre cifre. È meglio passare al logaritmo in base 10 di questi numeri: $y = log_{10}x$. Allora, $y = log_{10}m + n$. Mostriamo che, se una variabile casuale M su $[1, 10)$ segue la legge di Benford, allora la variabile casuale $Z = log_{10}M$ è semplicemente uniforme su $[0, 1)$. A tale scopo, è sufficiente mostrare che la funzione di distribuzione di $\mathbb{Z}$ è quella della variabile casuale uniforme su $[0, 1)$ cioè

$$F(z) = \begin{cases} 0, & z < 0, \\ z, & z \in [0, 1), \\ 1, & z \geq 1. \end{cases} \tag{4}$$

In effetti, quando $z \in [0, 1)$,

$$P(Z \leq z) = P(0 \leq log_{10}M \leq z) = P(1 \leq M \leq 10^z) = log_{10}10^z = z.$$

Se X appartiene all'insieme S_1, allora Y appartiene all'insieme $T_1 = log_{10}S_1$:

$$\begin{aligned} T_1 &= [0, log_{10}2) \cup [1, 1 + log_{10}2) \cup [2, 2 + log_{10}2) \\ &= \cup [3, 3 + log_{10}2) \cup [4, 4 + log_{10}2) \cup [5, 5 + log_{10}2), \end{aligned}$$

e analogamente per le altre cifre. Supponiamo che il prendere un numero casuale nel nostro insieme sia una variabile casuale X che assume valori nell'intervallo $[1, 10^6)$. Allora $Y = log_{10}X$ assume valori nell'intervallo $[0, 6)$. Si

ricordi che la probabilità che una qualche variabile casuale appartenga ad un qualche insieme è data dall'area sottesa al grafico della funzione di densità sull'insieme. Se la funzione di densità f di Y sull'insieme $[0, 6)$ fosse uniforme come in Figura 4(a), avremmo già concluso. Il più delle volte, comunque, ciò non si presenterà, come mostrato in Figura 4(b). **Ecco perché è così importante che l'insieme dei numeri originario comprenda diversi ordini di grandezza**. Le diverse porzioni corrispondenti ad una data prima cifra significativa i sono distribuiti orizzontalmente su diversi segmenti, la somma delle cui lunghezze è dell'ordine di $log_{10}(1 + \frac{1}{i})$ dell'ampiezza totale. Quindi, anche se l'altezza di $f(x)$ non è la stessa da un segmento all'altro, si potrebbe sperare che l'altezza media sia dello stesso ordine di grandezza per le varie cifre. Quando questo capita, allora i dati seguono la legge di Benford.

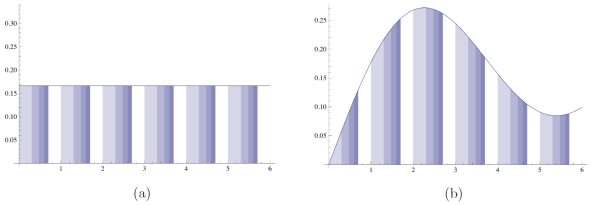


Figura 4: (a) la funzione di densità f uniforme, (b) la funzione di densità f non uniforme. In Figura sono rappresentate le aree corrispondenti alle frequenze delle prime cifre significative 1, 2, 3 e 4 per due differenti funzioni di densità di Y. I valori delle aree corrispondenti sono graficati in Figura 5.

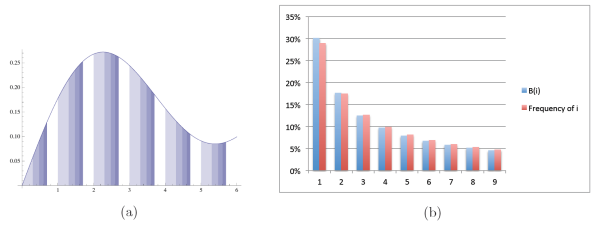


Figura 5: (a) la funzione di densità di f, (b) le aree sotto la curva delle cifre significative di f e per la funzione uniforme. In Figura sono rappresentate le aree corrispondenti alle frequenze delle prime cifre significative 1, 2, 3 e 4 per la funzione di densità di Figura 4(b). Sulla destra notiamo che questi stessi valori sono molto vicini a quelli ottenuti con la legge di Benford, nel caso di una funzione di densità uniforme per Y.

5. Come si verifica se un insieme di numeri segue la legge di Benford?

Se avete seguito un corso di statistica, allora avete probabilmente studiato il test del χ^2 per la bontà del campione. Questo test permette di testare se certi dati seguono una certa distribuzione di probabilità. Supponiamo che vogliate fare il test con un insieme di n numeri. Dovete solo costruire una tabella in cui n_i rappresenti il numero dei numeri del vostro insieme che hanno i come prima cifra significativa. Di certo, $n = n_1 + ... + n_9$. Gli N_i rappresentano i numeri di numeri che dovrebbero avere i come prima cifra significativa se il vostro insieme seguisse la legge di Benford, cioè $N_i = nB(i)$.

i	1	2	3	4	5	6	7	8	9
n_i	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9
N_i	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8	N_9

Figura 6: La tabella per il test del χ^2 per la bontà del campione

Quindi, si calcola

$$\chi^2 = \sum_{i=1}^{9} \frac{(n_i - N_i)^2}{N_i}$$

e si guarda in una tabella del χ^2 nella linea corrispondente a 8 gradi di libertà. Se si vuole fare il test con il 5% di errore, allora, se $\chi^2 < 15.51$, si accetta che i dati soddisfino la legge di Benford, altrimenti lo si rifiuta. Questa è una ricetta veloce, ma per fare questo tipo di test con gli studenti occorre spendere del tempo per familiarizzare coi dettagli del test e il suo significato.

6. Invarianza della legge di Benford per cambiamenti di base

Tale invarianza potrebbe essere modellizzata in modo analogo a quella per cambiamenti di scala. Tuttavia è una questione più delicata, perché non possiamo limitare il nostro lavoro alla mantissa. Infatti, se $x = m10^n$, allora anche la parte 10^n deve essere convertita nella nuova base. E, in effetti, la principale difficoltà è esprimere in termini matematici che cosa significa per una variabile casuale essere indipendente per cambiamenti di base. Tralasceremo i dettagli.

7. Conclusione

La legge di Benford è affascinante: essa sfida l'intuizione, ed è qualcosa che si può testare da sé e persino adattare ad un'attività di classe. È usata per

essere una curiosità , ma ora è uno strumento standard per intercettare le frodi. Sicuramente, sempre più evasori fiscali la conoscono. Ma prestate attenzione: la prima cifra significativa non è la sola cosa di cui tenere conto. La legge di Benford generalizzata permette di ottenere una legge per la seconda cifra significativa, per la terza, ecc... Potete provare a trovarla da soli: basta pensare in quali unioni di intervalli la mantissa di un numero dovrebbe essere tale da avere i come seconda cifra significativa.