Pandemia da Covid 19 e Legge di Benford

Filippo Elba - 01/10/2020 [papers]

Abstract

È qui proposta un'applicazione della legge di Benford (o della prima cifra) ai dati giornalieri per Paese dei nuovi casi registrati di contagi da Covid 19. La fonte dati utilizzati è https://ourworldindata.org/coronavirus-source-data. L'intento principale di questo scritto è quello di verificare se la distribuzione delle prime cifre dei numeri relativi ai casi giornalieri registrati rispetta la distribuzione della legge di Benford. Seguono alcune brevi osservazioni sui risultati ottenuti.

Legge di Benford: caratteristiche

La legge di Benford (o legge della prima cifra) è una distribuzione di probabilità che descrive la probabilità che un numero, presente in una raccolta di dati reali (popolazione dei comuni, quotazione delle azioni, costanti fisiche o matematiche, etc.), cominci con una data cifra. Tale funzione di probabilità è uguale a: (1)[1]

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}\left(1+\frac{1}{d}\right) \qquad d = 1, 2, \dots, 9$$

La particolarità di quanto affermato da questa legge sta nel fatto che, stanti alcune condizioni, le probabilità che un numero *x* avrebbe di iniziare con una cifra compresa tra 1 e 9 non sarebbero tra loro uguali (ossia dello 0,11 ca.).

Negli anni '30, Benford[2] dimostra ciò ricorrendo a più di 20.000 casi concernenti le più disparate serie di dati. In ogni verifica da lui effettuata, le frequenze con cui i numeri degli insiemi presi in esame iniziano con la cifra "1" sono di gran lunga maggiori rispetto a tutte le altre e sempre vicine alla quota di un terzo rispetto al totale. Fenomeno che non sembra riscontrarsi quando l'analisi passa alle cifre successive. In questi casi, infatti, le frequenze relative sono più uniformi rispetto alla probabilità dello 0,10 (dalla seconda cifra in avanti c'è da aggiungere la possibilità che essa possa assumere valore "0").

Per le cifre successive alla prima varrebbe la seguente funzione di probabilità: (2) [3]

$$P(d) = \sum_{k=10^{n-2}}^{10^{n-1}-1} \log_{10}\left(1+\frac{1}{10k+d}\right) \quad con\ n > 1;\ \ d = 0, 1, \dots, 9$$

Tale legge sarebbe indifferente rispetto alla scala di misurazione adottata. Dimostrazioni effettuate da Benford e da altri confermano questa proprietà[4]. Purché le unità di misura adottate non siano espresse in base al sistema binario, la maggior frequenza con cui si registrano numeri con la prima cifra piccola sarebbe sempre empiricamente confermata.

Affinché la frequenza delle prime cifre di dati riferiti a un qualsiasi insieme siano quanto più vicini possibili alle probabilità benfordiane, è determinante il rispetto di alcuni "paletti":

- l'insieme dei numeri deve essere scelto sulla base di una variabile casuale;
- tutte le cifre da 1 a 9 devono avere stessa probabilità di poter essere al primo posto nei numeri che costituiscono l'insieme, senza alcun limite, anche inconsapevole;
- i numeri dell'insieme devono essere molti, evitando ripetizioni;

- l'insieme non deve essere costituito da numeri "assegnati" (per esempio, i numeri telefonici);
- il valore della media dei numeri dell'insieme deve essere maggiore rispetto alla mediana;
- i numeri dell'insieme considerato devono essere costituiti da più cifre (meglio se almeno quattro[5]).

È in particolare Hill (1995)[6] che si interessa alla determinazione delle caratteristiche che deve avere un insieme di numeri per poter meglio soddisfare le probabilità di distribuzione benfordiane: se si scelgono in modo casuale delle distribuzioni che godono della proprietà di essere invarianti rispetto alla scala di misurazionee dei campioni casuali siano presi da ciascuno di essi, le frequenze relative delle prime cifre dei numeri così ottenuti seguiranno, molto probabilmente, la legge di Benford. Sarebbero le distribuzioni di seconda generazione, quelle risultato della combinazione di altre distribuzioni, ad adattarsi meglio alle previsioni benfordiane.

Come già precedentemente sottolineato, quella di Benford è una legge empirica e, in quanto tale, difficile è cercarne un fondamento teorico. In *The Law of Anomalous Numbers* (1937), tuttavia, Benford rileva come molti fenomeni in natura si caratterizzerebbero per il fatto di seguire scale logaritmiche o geometriche piuttosto che aritmetiche. Tra gli esempi che egli cita: la crescita della sensibilità della retina alla brillantezza all'aumentare progressivo dell'illuminazione, dell'orecchio all'aumentare della rumorosità, la cognizione del tempo all'aumentare dell'età. In generale, negli ambiti più diversi fra loro (dalla medicina alla fisica atomica), tenderebbe sempre a presentarsi questa situazione, negando, perciò, la "naturalezza" del concetto di proporzionalità costante. Come lo stesso Bendford afferma alla fine del suo scritto "*small things are more numerous than large things, and there is a tendency for the step between sizes to be equal to a fixed fraction of the last preceding phenomenon or event*"[7].

Alcuni casi pratici di applicazione della legge

Diversi sono gli ambiti in cui si è cercato di applicare la legge di Benford: da indagini riguardanti presunti brogli elettorali, a controlli effettuati sui mercati finanziari durante la crisi del 2007-08[8] o relativi alla veridicità dei risultati pubblicati nelle riviste scientifiche[9]. Tra le applicazioni più note della legge di Benford ci sono quelle di Varian e di Nigrini.

Nel 1972, Varian suggerí la possibilità di utilizzare questa legge per individuare eventuali falsificazioni nelle raccolte di dati usate per supportare decisioni politiche, basandosi sul presupposto che chi vuole addomesticare i dati sarebbe portato ad usare numeri distribuiti in modo non naturale, in contrasto, dunque, con le previsioni benfordiane. Comparando la frequenza relativa delle prime cifre di un insieme di numeri con le frequenze di Benford, si potrebbero evidenziare risultati anomali. Alla stessa maniera si può usare questo confronto per cercare falsificazioni in raccolte di dati riguardanti, per esempio, costi e entrate.

Altra applicazione conosciuta è quella fatta da Nigrini (1996)[10]. Dopo aver provato l'efficacia della legge di Benford su casi reali di frode accertata riferiti al 1992, egli studia come utilizzarla in maniera sistematica per testare il contenuto delle dichiarazioni dei redditi. Un caso peculiare, spesso ricordato dallo studioso americano per confermare la validità dell'utilizzo di questa legge nell'ambito dei controlli fiscali, si verifica nel 1993. In quell'anno, lo stato dell'Arizona cita in giudizio W. J. Nelson, accusato di aver cercato di defraudare lo stato per circa due milioni di dollari[11]. Nelson, manager dell'ufficio del Tesoro dell'Arizona, respinge le accuse dimostrando che non vi è alcun tipo di prova che possa confermare ciò. In realtà, le ventitré operazioni sospette imputategli registrano, il più delle volte, un ammontare di poco inferiore ai cento mila dollari (soglia passata la quale sarebbero previsti controlli aggiuntivi da parte di organismi appositi). Ciò implica un aumento del numero di volte in cui il "7", l'"8" e il "9" rappresentano la cifra iniziale delle operazioni effettuate da questo funzionario (90% il loro totale, in netta controtendenza rispetto a quanto previsto dalla legge della prima cifra).

Legge di Benford sui dati dei contagi[12]

Il presente lavoro presenta un'applicazione della legge di Benford ai numeri dei contagi giornalieri registrati in un gruppo di Paesi (Brasile, Cina, Francia, Germania, Giappone, Italia, Regno Unito, Russia, Sud Corea e Stai Uniti)[13]. La rilevazione della banca dati utilizzata parte dal 31 dicembre 2019. Il calcolo delle frequenze per ogni Paese tiene conto dei soli giorni in cui si è registrato un numero di casi maggiore di zero. È per questo che per alcuni Paesi, vedi Cina, si dispone di quasi duecentoquaranta osservazioni. Per altri, più recentemente toccati dal contagio, di poco più di centottanta (per esempio, Brasile e Francia).

Prima di confrontare le distribuzioni delle "prime cifre", è utile una breve panoramica sull'andamento dei casi giornalieri registrati nei Paesi che sono oggetto della presente analisi.

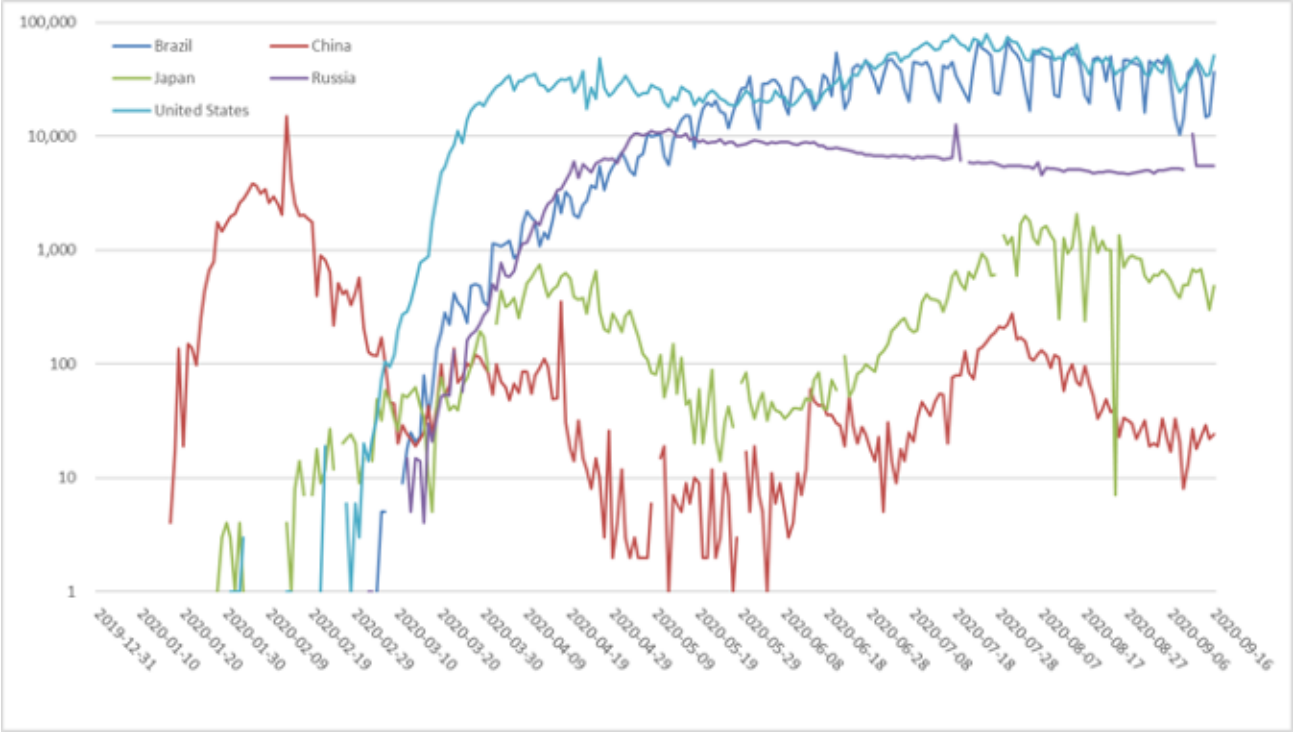
Tra i Paesi esaminati con una popolazione oltre i 100 milioni di abitanti (Figura 1a), la Cina, primo Paese in assoluto a registrate casi di Covid 19, raggiunge un picco massimo a metà febbraio (i primi contagi si registrano circa un mese prima). Dopo una lunga decrescita che porta a quasi azzerare i casi tra metà maggio e i primi di giugno, affronta un nuovo incremento nel corso dell'estate, con numeri comunque molto al di sotto dei picchi invernali. Attualmente la tendenza è in discesa.

Brasile e Stati Uniti seguono un andamento molto simile tra loro: primi casi registrati tra fine febbraio e i primi di marzo, successivo andamento crescente che, seppur attenuato, sembra persistere fino a luglio. Nell'ultimo periodo si evidenzia, invece, un lieve calo.

La Russia, dopo aver raggiunti i livelli massimi verso fine maggio (circa 10 mila nuovi casi giornalieri), conosce un lieve decremento nel corso di tutta l'estate. Situazione più altalenante per il Giappone il quale, specie da aprile in poi, sembra seguire l'andamento cinese ma con un numero più elevato di casi: primo picco raggiunto nei primi di aprile, poi decremento e nuovo picco, più alto di quello primaverile, attorno alla metà di agosto. Nelle ultime settimane tendenziale rallentamento.

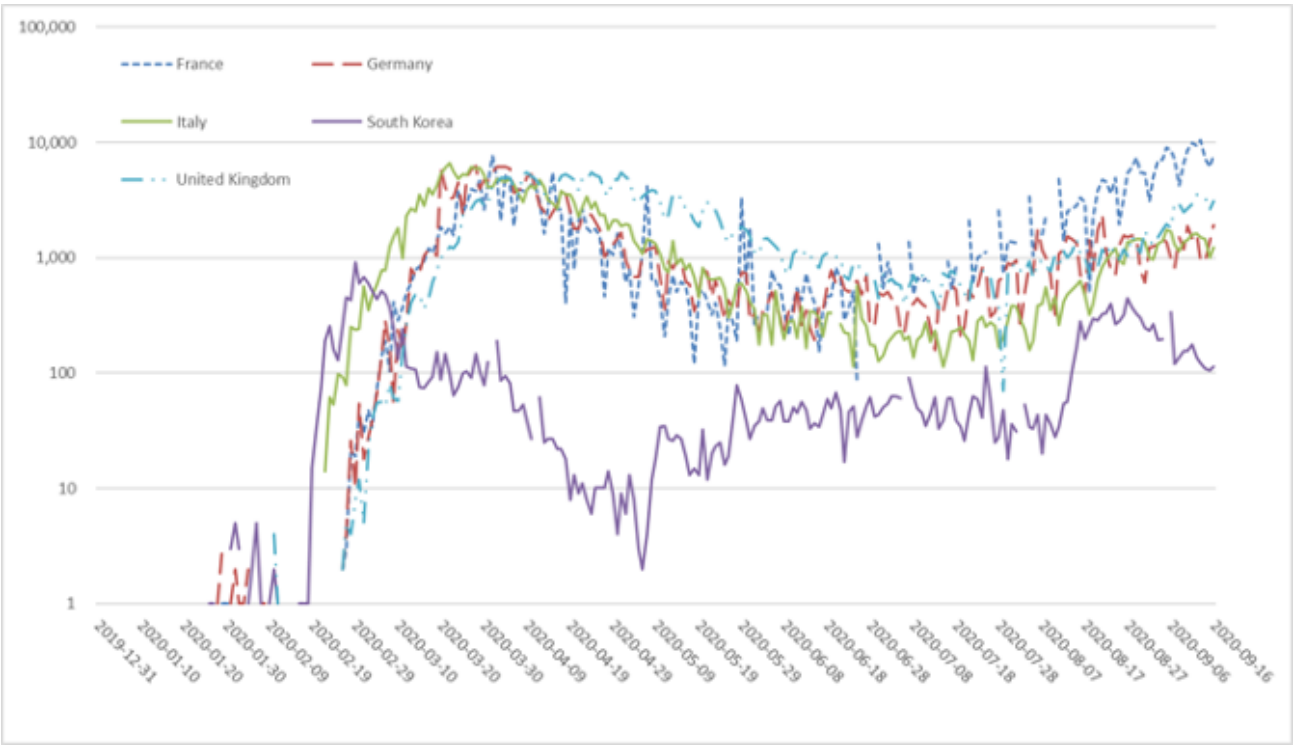
Alcuni dei Paesi considerati tra quelli con meno di 100 milioni di abitanti (Figura 1b) registrano i primi casi già alla fine di gennaio, ma in maniera più seria tra febbraio e marzo. La Corea del Sud raggiunge il proprio picco già a fine febbraio, i Paesi europei attorno alla prima metà di aprile. Segue per tutti un generalizzato calo fino ai primi di maggio per la Corea del Sud, metà giugno per i Paesi europei, poi nuovi incrementi. Nelle ultime settimane, fatta eccezione per il Paese asiatico che conosce un nuovo decremento, gli altri affrontano un rialzo, con superamento, nel caso francese, del picco primaverile.

1a. Nuovi casi giornalieri (scala log10): Paesi con >100 mln ab.



Fonte: elaborazioni su dati https://ourworldindata.org/coronavirus-source-data.

1b. Nuovi casi giornalieri (scala log10): Paesi con <100 mln ab.



Fonte: elaborazioni su dati https://ourworldindata.org/coronavirus-source-data.

Passando adesso al confronto delle distribuzioni delle prime cifre dei numeri dei casi giornalieri registrati nei Paesi del gruppo con la distribuzione di Benford (Figure 2 e 3), utile è valutare i risultati di un test chi-quadro tra la distribuzione effettivamente

osservata e quella di riferimento di Benford. Si effettua anche un confronto tra la prima e una distribuzione uniforme, ipotizzando che la probabilità con cui una delle nove cifre si presenti al primo posto nel numero di nuovi casi giornalieri registrati sia pari allo 0.11.

Si nota, innanzitutto, come le distribuzioni della prima cifra di tutti gli otto Paesi siano statisticamente differenti da quella uniforme (Tabella 1, lato sinistro).

Il test del chi quadro che confronta invece le distribuzioni osservate con quella di Benford (Tabella 1, lato destro) conferma come, in alcuni casi (cinque su dieci), non vi sarebbe una differenza statisticamente significativa tra le distribuzioni.

1. Test chi-quadro: confronto distribuzioni prime cifre Paesi vs distribuzione discreta uniforme (sx) e distribuzione Benford (dx)

	Nr. oss.	p-value	Sign. 95%	Sign. 99%
Brasile	189	0.0000		
Cina	238	0.0000		
Francia	186	0.0000		
Germania	207	0.0000		
Giappone	214	0.0000		
Italia	201	0.0000		
Regno Unito	204	0.0000		
Russia	184	0.0000		
Stati Uniti	210	0.0000		
Sud Corea	216	0.0000		

	Nr. oss.	p-value	Sign. 95%	Sign. 99%
Brasile	189	0.0000		
Cina	238	0.9331	*	*
Francia	186	0.0090		
Germania	207	0.2237	*	*
Giappone	214	0.0840	*	*
Italia	201	0.0751	*	*
Regno Unito	204	0.0038		
Russia	184	0.0000		
Stati Uniti	210	0.0000		
Sud Corea	216	0.0860	*	*

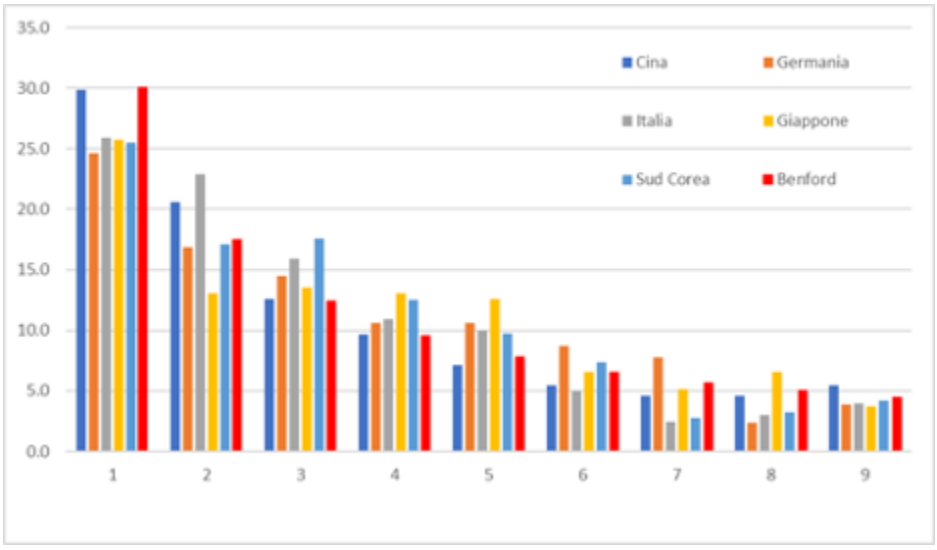
Fonte: elaborazioni su dati https://ourworldindata.org/coronavirus-source-data.

Partendo proprio dalla distinzione tra Paesi con distribuzione significativamente diversa e non diversa da quella di Benford, è di seguito presentata un'analisi basata su confronto grafico.

Per i Paesi che registrano una distribuzione "più simile" a quella di Benford in base al test del chi-quadro (Figura 2), si nota una notevole somiglianza. La Cina in particolar modo (vedi cifre 1, 3, 4 su tutte). È probabilmente il Giappone che sembra discostarsi maggiormente (vedi cifre 4, 5, 8).

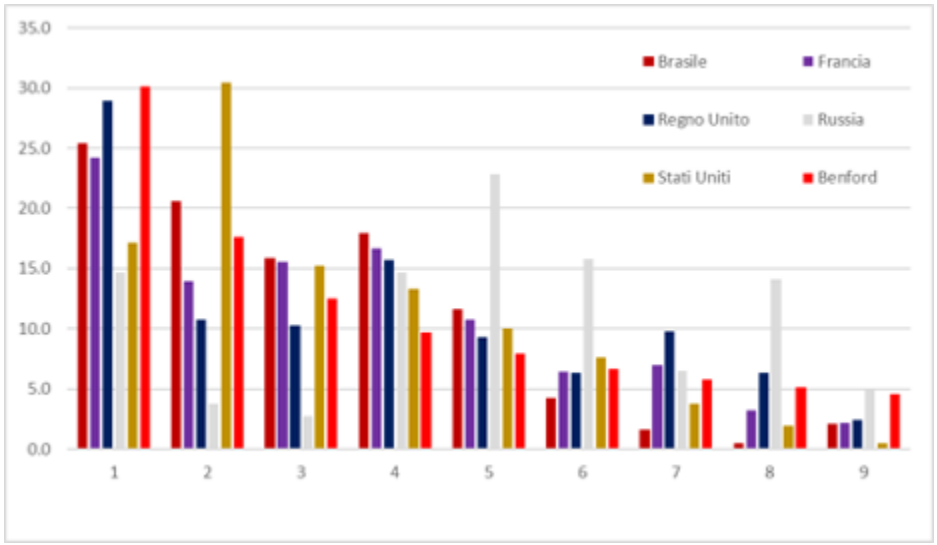
Per i Paesi che in base al test chi-quadro avrebbero una distribuzione delle prime cifre significativamente diversa da quella di Benford (Figura 3), si evidenzia come, tuttavia, esista una somiglianza di fondo, e cioè che le cifre più piccole si presentano tendenzialmente con maggior frequenza rispetto a quelle più grandi. Forse l'unica eccezione è rappresentata dalla Russia, la cui distribuzione non sembra seguire quella di Benford neanche lontanamente.

2. Paesi con distr. prima cifra non significativamente diversa da legge Benford



Fonte: elaborazioni su dati https://ourworldindata.org/coronavirus-source-data.

2. Paesi con distr. prima cifra significativamente diversa da legge Benford



Fonte: elaborazioni su dati https://ourworldindata.org/coronavirus-source-data.

Osservazioni conclusive

L'analisi qui condotta dimostra come questi dati sembrino rispettare la distribuzione delle prime cifre della legge di Benford. Anzi, per alcuni dei Paesi considerati, oltre all'evidenza grafica, vi sarebbe anche una evidenza statistica. Anche in questo caso, dunque, sembrerebbe essere confermato il contenuto di questa affascinante legge empirica.

**Questo lavoro è l'ampliamento di un'analisi già realizzata nel mese di luglio 2020 su un numero più piccolo di Paesi:* https://www.researchgate.net/publication/343448433_PANDEMIA_DA_COVID_19_E_LEGGE_DI_BENFORD .

Bibliografia

Benford F. (1937), *The law of anomalous numbers*, Proceedings of the American Philosophical Society, 78(4).

Diekmann A. (2004), *Not the first digit! Using Benford's law to detect fraudolet scientific data*, Journal of applied statistics (http://128.118.178.162/eps/othr/papers/0507/0507001.pdf).

Durtschi C., Hillison W. e Pacini C. (2004), *The effective use of Benford's law to assist in detecting fraud in accounting data*, Journal of Forensic Accounting vol. V, pagg. 17-34 (http://www.auditnet.org/articles/JFA-V-1-17-34.pdf).

Elba F. (2013), *Legge di Benford nella lotta all'evasione fiscale*, unpublished working paper.

Gunnel S. e Todter K.H. (2007), *Does Benford's Law hold in economic research and forecasting?*, Discussion paper: economic studies, Deutsche Bundesbank.

Hill T. (1995), *A statistical derivation of the Significant-Digit law,* Statistical Science 10 (4), pagg. 354-363.

Hofmarcher P. e Hornik K. (2010), *First Significant Digits and the Credit Derivative Market during the Financial Crisis*, Research Report Series / Department of Statistics and Mathematics, 101. Institute for Statistics and Mathematics, WU Vienna University of Economics and Business, Vienna.

Ley E. (1996), *On the peculiar distribution of the U.S. stock indexes' digits,* The American Statistician, vol. 50 n.4, pagg. 311-313.

Ley E. e Varian H. (1994), *Are there psychological barriers in the Dow-Jones Index?*, Applied Finacial Economics, n.4, pagg. 217-224.

MathWord – Benford's Law (http://mathworld.wolfram.com/BenfordsLaw.html).

Mittermaier L. e Nigrini M. (1997), *The use of Benford's law as an aid in analytical procedures,* A Journal of Practice & Theory, vol.16, n.2, pagg. 52-67.

Nigrini M. (1996), *A Taxpayer Compliance Application of Benford's law,* Journal of the American Taxation Association 18, pagg. 72-91.

Nigrini M. (1999), *I've got your number*, Journal of Accountancy (http://www.journalofaccountancy.com/issues/1999/may/nigrini.htm).

Offreddi M. (2010), *La legge di Benford per l'analisi statistica dei bilanci aziendali*, tesi di laurea a.a. 2009/2010, Università degli Studi di Brescia.

Schiavo M. (2010), *Analisi statistica sulla legge di Benford*, tesi di laurea a.a. 2009/2010, Università degli Studi di Padova.

Rosario M. R. e Zizza R. (2008), *L'evasione dell'Irpef: una stima per tipologia di contribuente,* mimeo, Banca d'Italia.

Varian H. (1972), *Benford's law,* The American Statistician 26 (3), pagg. 62-66.

---

[1]La (1) è applicabile anche a stringhe di cifre. In questo caso *d* assume il valore della stringa d'interesse.

[2]In realtà, già prima di Benford, pare che questa legge fosse stata scoperta, dalla fine del XIX secolo, dal matematico e astronomo Simon Newcomb. Egli notò che, le pagine dei libri con le tabelle dei logaritmi riferite a numeri aventi "1" come prima cifra, erano più sporche delle altre. Ne dedusse che, probabilmente, ciò dipendesse dal fatto che venivano usate più spesso. Venne controargomentato, tuttavia, che in qualsiasi libro, al quale si accede alle pagine in modo sequenziale, le prime tra esse

sarebbero state sempre più usate delle ultime.

[3]Con *n* che rappresenta la posizione occupata dalla cifra *d* nella stringa numerica (sui valori assunti delle frequenze relative riferite alla seconda cifra in base alla (2): Tab.2).

[4]MathWord – Benford's Law (http://mathworld.wolfram.com/BenfordsLaw.html).

[5]Affinché sia rispettata al meglio la distribuzione benfordiana sulla prima cifra, ideale sarebbe considerare numeri dell'ordine delle migliaia. In questo modo la variabile "prima cifra del numero" si configurerebbe come continua e non come discreta (cosa che accade, in particolare, per i numeri costituiti da unica cifra). Per una miglior descrizione di quanto detto: Benford F. (1937), *The Law of Anomalous Numbers,* Proceedings of the American Philosophical Society 78 (4), pagg. 551–572.

[6]Hill T. (1995), *A statistical derivation of the Significant-Digit Law,* Statistical Science 10 (4), pagg. 354-363.

[7]Benford F. (1937), *op. cit.*, pag. 571.

[8]Hofmarker P. e Hornik K. (2010), *First significant digit and the credit derivative market during the financial crisis,* Research Report Series / Department of Statistics and Mathematics,101. Institute for Statistics and Mathematics, WU Vienna University of Economics and Business, Vienna.

Diekmann A. (2004), *Not the first digit! Using Benford's Law to detect fraudolet scientific data*, Journal of applied statistics (http://128.118.178.162/eps/othr/papers/0507/0507001.pdf).

[10]Nigrini M. (1996), *A Taxpayer Compliance Application of Benford's Law.,* J. Amer. Tax. Assoc. 18, 72-91, 1996.

[11]Nigrini M. (1999), *I've got your number*, Journal of Accountancy. (http://www.journalofaccountancy.com/issues/1999/may/nigrini.htm).

[12] Il software utilizzato per le analisi è Microsoft Excel 2020.

[13] Esiste una lunga lista di articoli e blog che realizzano la stessa analisi. Allo stato attuale, però, nessuna in lingua italiana.