UNIVERSITÀ DEGLI STUDI DI FERRARA

Benford's Law: genesi, letteratura e applicazioni empiriche

Yannick Tazzari, Stefano Bonnini, Giuseppe Marzo

Abstract

Data analysis and detection of anomalies

Benford's Law is the statistical law capable of identifying the hidden anomalies of large datasets. The purpose of this article is to summarize how this logarithmic law was discovered, to summarize the related literature and to describe its empirical applications.

1. Introduzione

La Legge di Benford, o legge della prima cifra significativa (*first-digit*), afferma che se si calcolano le frequenze delle prime cifre significative in una collezione di numeri distribuita in maniera casuale, i numeri con 1 come prima cifra significativa dovrebbero apparire circa il 30% delle volte, mentre i numeri con 2 come prima cifra significativa appaiono circa il 17% delle volte e così a scalare fino alla cifra 9 che compare circa il 4,5% delle volte.

Se i *dataset*, conformi a questa legge, sono manipolati o presentano anomalie allora con l'applicazione dell'analisi di Benford è possibile individuare tali manipolazioni.

2. Benford's Law

La Legge di Benford misura e descrive 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. Ad esempio, la prima cifra significativa di 3.547 è 3 e quella di 0.00732 è 7.

Diversamente, si può definire tale legge statistica scrivendo un numero reale positivo x come un numero $m \in [1,9)$ moltiplicato per una potenza di 10: $x = m10^n$, $n \in \mathbb{Z}$.

La prima cifra significativa di x è la parte intera di m che può essere denotata con $[m]$.

Il numero m è detto mantissa di x.

Se si collezionano numeri in modo casuale e si calcola la frequenza $B(i)$ della prima cifra significativa di i, allora $B(i)$ è data approssimativamente da $\log_{10}(1+\frac{1}{i})$.

Ne deriva che le frequenze sono le seguenti:

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

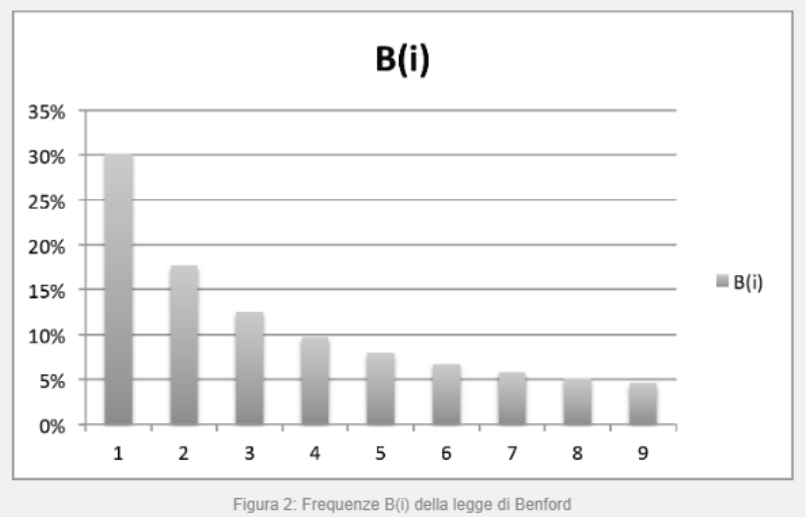

Figura 2: Frequenze B(i) della legge di Benford

Figura 1 - Distribuzione della prima cifra significativa

3. La genesi di Benford's Law

Si fornisce una breve nota storica di come è stata scoperta la legge di Benford's Law (di seguito anche "BL") oggetto del presente elaborato.

Nel 1881 l'astronomo Newcombe pubblica l'articolo intitolato "*Note on the Frequency of Use of the Different Digits in Natural Numbers*"[1], nel quale descrive la propria scoperta.

Egli si rende conto, nell'osservare le pagine delle tavole logaritmiche (non solo il contenuto, ma anche il numero di pagina) che i bordi dei primi fogli (con le tavole aventi "1" come prima cifra) sono maggiormente usurati rispetto a quelli degli ultimi, presumibilmente, poiché sfogliate, quindi utilizzate, più frequentemente.

Da questa riflessione, Newcombe inizia a formulare il principio che, in ogni lista di numeri ricavati da un insieme arbitrario di dati, la maggior parte di essi tende a cominciare con "1" rispetto a qualsiasi altra cifra. Al procedere delle cifre si assiste ad un regredire della frequenza con cui queste appaiono come *first-digit*.

Questa osservazione è in contrasto con la cultura del periodo storico in cui l'astronomo vive, secondo la quale *"In una sequenza di numeri casuali la probabilità che uno di essi inizi per 1 o per 9 è praticamente la stessa*." Infatti la scoperta di Newcombe, essendo priva del supporto di applicazioni empiriche, non attira la dovuta attenzione per diversi anni.

Quasi un mezzo secolo dopo, presso i laboratori della General Electric Company, il fisico di nome Frank Benford, ignaro dei precedenti lavori di Newcombe, esprime la medesima intuizione e formula una riflessione similare a quella del precursore[2]. A differenza del primo però, egli esegue applicazioni empiriche della Legge, ottenendo conferme dalle osservazioni dirette e dalle riflessioni compiute su di esse. Al fine di convalidare la propria assunzione, Benford raccoglie una collezione di 20.000 dati numerici, di differente origine e variegata natura, relativi a ventuno elementi tra di loro assai diversi, misurando la frequenza delle cifre da 1 a 9, compresi, escludendo dall'analisi la cifra 0.

Nel 1938, Benford rende pubbliche le proprie ricerche enunciando *The Law of Anomalous Numbers*[3] che, negli anni a seguire, viene ribattezzata da altri studiosi con il nome del proprio creatore: Benford's Law.

[1]Newcombe Simon, Note on the Frequency of Use of the Different Digits in Natural Numbers, "*American Journal of Mathematics"*, Vol. 4, No. 1. (1881), pp. 39-40.

[2]Hassan, B. (2002). Assessing data authencity with Benford's law. Information Systems Control Journal 6.

[3]Benford Frank, (1938), The law of anomalous numbers, Physicist, Research Laboratory, General Electric Company, Schenectadt, New York, *American Philosophical Society.*

4. L'approccio di analisi di Frank Benford

Nella propria pubblicazione, F. Benford analizza il principio della legge logaritmica ed espone i risultati dello studio effettuato su numerose successioni numeriche. Nonostante queste ultime mostrino evidenza della suddetta legge logaritmica, permane l'esigenza di trovare una giustificazione della stessa. L'articolo di Benford si sviluppa sui seguenti diversi punti.

- PARTE 1, *Statistical Derivation of the Law*, che si occupa di definire i seguenti principi:

  I. I metodi e i termini d'uso del principio:

  - Si distingue tra la cifra e il numero, quest'ultimo si compone di uno o più cifre, e può avere 0 come cifra in una qualsiasi posizione dopo la prima;

  - Il metodo di studio consiste nel selezionare qualsiasi *dataset* che non abbia un limite imposto nell'intervallo numerico;

  - Si conta il numero di volte in cui i numeri naturali 1, 2, 3, ... 9 si verificano come prime cifre del numero;

  - Se un punto decimale o pari a 0 si verifica davanti al primo numero naturale, esso deve essere ignorato, poiché nessuna attenzione deve essere rivolta a grandezze differenti da quella indicata dalla prima cifra significativa.

  II. La legge dei grandi numeri:

  Occorre raccogliere dati da quanti più campi di ricerca e settori possibili così da rendere il campione sufficientemente ampio.

  III. La frequenza delle cifre nella prima posizione:

  L'analisi dei *dataset* consiste nel misurare la percentuale di volte in cui ciascuno dei numeri naturali da 1 a 9 è utilizzato come prima cifra nei numeri, ovvero la percentuale media di distribuzione di ogni cifra e il relativo scostamento percentuale.

  IV. I reciproci:

  Nell'analisi dei *dataset* è opportuno considerare che vi sono alcuni tabulati di dati tecnici e scientifici che sono forniti in forma *reciproca*, si pensi ad esempio a candela per watt e a watt per candela. Se la forma di una tabulazione segue una distribuzione logaritmica, allora la tabulazione reciproca deve avere la stessa distribuzione.

  V. La legge dei numeri anomali:

  Si riscontra maggiore aderenza alla Legge Logaritmica per i dati aventi natura casuale piuttosto che per quelli aventi natura formale o matematica. In particolar modo, si verifica

che l'insieme di numeri arabi (non enunciati) di numeri presenti in pagine consecutive di un libro sono conformi alla BL, così come il *dataset* selezionato dall'autore contenente i primi 342 indirizzi stradali degli studiosi di scienza americani.

- PARTE 2, *Geometric Basis of the Law*, che si focalizza su:

  I. Le serie geometriche e logaritmiche:

  Sussiste una stretta relazione tra la serie geometrica e la serie logaritmica se si considera la curva tracciata dalla legge logaritmica.

  II. Le curve semi-log:

  L'analisi del *dataset* composto dagli indirizzi stradali dei primi 342 studiosi di scienza americani selezionato da Benford, e la relativa distribuzione, soddisfano la relazione logaritmica e rappresentano le quattro serie geometriche raffigurate nel seguente grafico.

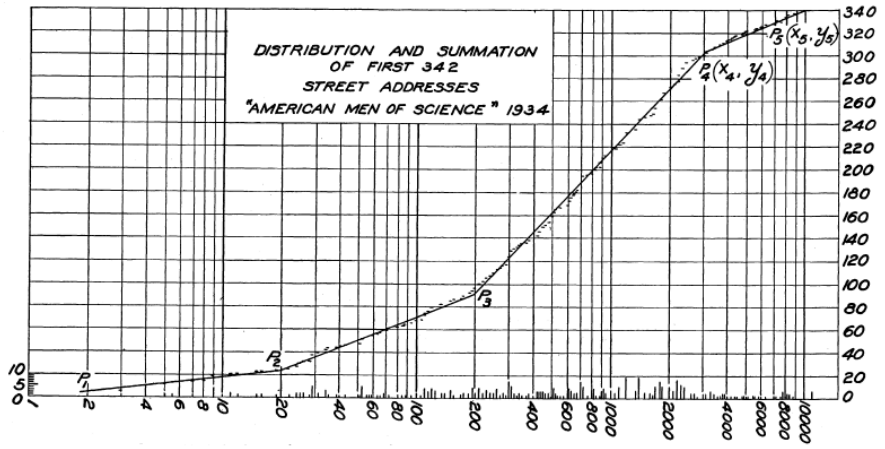

Figura 2 - Curva semi-log

  III. I numeri naturali e la natura dei numeri:

  Occorre considerare che negli eventi naturali ci sono molti esempi di progressioni geometriche o logaritmiche. Infatti, nonostante si abbia la propensione ad etichettare qualsiasi oggetto con una progressione numerica del tipo 1, 2, 3, 4, * * *, e a reputare la sequenza 1, 2, 4, 8, * * * una disposizione naturale non accettabile, anche quest'ultima rappresenta un ordine con cui i fenomeni possono verificarsi. Si consideri, per esempio, alla reazione fisiologica e psicologica a stimoli esterni nel campo della medicina, nonché alla crescita della sensazione di luminosità all'aumentare dell'illuminazione, al "senso del volume" ed al "senso del peso".

- PARTE 3*, Digital Order of Numbers,* che si concentra su:

  I. La frequenza teorica in vari ordini di cifre;

II. La somma delle frequenze.

5. Letteratura e applicazioni empiriche di Benford's Law

Negli anni successivi alla formulazione ed enunciazione di Benford's Law del 1938 vi sono interventi e contributi da parte di studiosi provenienti da molteplici campi di ricerca, ossia matematici, statistici, economisti, ingegneri e fisici. Grazie a questi apporti si individuano le intrinseche proprietà della legge statistica, ovvero la scala invarianza e la base invarianza e altre caratteristiche ad essa legate, quali:

- le condizioni e i presupposti di conformità dei *dataset*,
- le regole delle assunzioni distributive,
- la struttura con cui eseguire l'analisi sui dati (Software),
- le operazioni che consentono di avere maggiore aderenza a BL,
- i test di valutazione della bontà di adattamento dei *dataset* a BL.

Benford stesso, come detto nel paragrafo precedente, cerca di spiegare il fenomeno della Legge Logaritmica effettuando analisi su collezioni di numeri interi naturali, nel tentativo di fornire prova che tale manifestazione si realizza naturalmente nel classico sistema di numerazione.

Tuttavia, Benford incontra alcuni problemi nel fornire spiegazione al fenomeno dovuti al fatto che questa serie non ha alcuna frequenza naturale asintotica. Tanto è vero che se si volesse rappresentare il comportamento del fenomeno, estrapolandolo da una sequenza di numeri interi, si otterrebbe questo grafico:

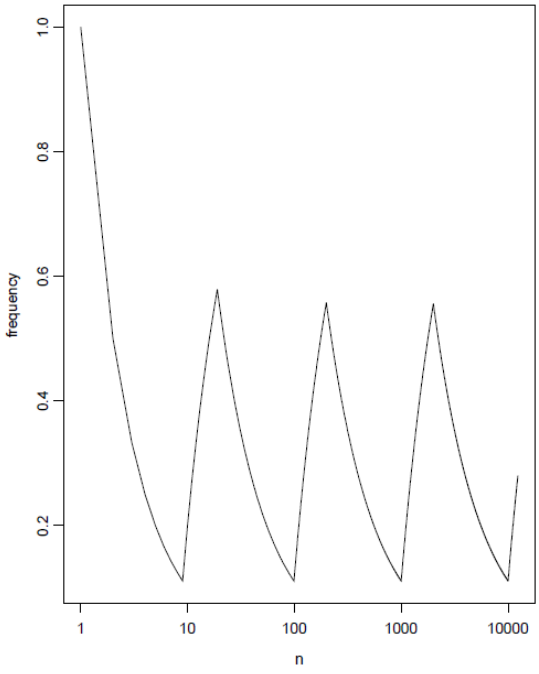

Figura 3 - Aspetto della frequenza di 1 come prima cifra significativa in una sequenza {1, … n}

Siccome il limite non esiste, la sequenza oscilla su una scala logaritmica tra due estremi non costanti, tale per cui:

- i minimi decrescono verso 1/9 = 0.1111 e,
- i massimi variano intorno a 5/9 = 0.5555.

Anche i successori di Benford cercano una spiegazione logica del fenomeno, ma senza riuscirvi perché il problema persiste per la ragione individuata da Hill nel 1995: alcuni *dataset* di dati di numeri "naturali" non aderiscono a BL .

Si presentano in ordine cronologico i principali contributi di studiosi e ricercatori che si sono dedicati a BL.

- ***1944 - 1994***
  - ***Roger S. Pinkham[4]***

Una delle prime spiegazioni dei risultati ottenuti da Benford è accreditata a Roger S. Pinkham, che nel 1961 comprova che la BL è

- valida indipendentemente dalle unità di misure adottate, cioè è *invariante di scala,*
- l'unica Legge a godere di tale proprietà.

L'autore dimostra empiricamente la sua intuizione osservando la lunghezza dei fiumi americani in miglia e in chilometri, ottenendo come risultato numeri diversi ma identiche frequenze relative alla prima cifra significativa.

  - ***Knuth D. E.[5]***

Altre prove empiriche supplementari a quelle dello stesso Benford sono condotte dal fisico Knuth, che nel 1969 osserva che le costanti fisiche più comunemente utilizzate (ad esempio, le costanti come la velocità della luce e la forza di gravità) seguono la Legge in quanto circa il 30% dei numeri ha come prima cifra significativa l'1.

  - ***Varian, H.R.[6]***

L'autore nel 1971 suggerisce la possibilità di utilizzare BL per individuare eventuali falsificazioni nelle raccolte di dati utilizzate per il supporto alle decisioni politiche. Al fine di rintracciare potenziali risultati anomali in ambito politico, l'autore propone di confrontare la frequenza relativa della prima cifra dei numeri utilizzati per il supporto a decisioni politiche con quelle "teoriche" di BL. Nel

[4]Pinkham, R.S. (1961). On the distribution of first significant digits. *Annals of Mathematical Statistics* 32, 1223-1230.
[5]Knuth, D. (1969). The Art of Computer Programming, vol. 2, 219-229. (2nd ed.). Addison-Wesley, Reading, MA, 239-249. (3rd ed.) (1998), 254-262.
[6]Varian, Hal (1972). "Benford's Law (Letters to the Editor)". *The American Statistician* 26 (3): 65. doi:10.1080/00031305.1972.10478934.

condurre questa ricerca, Varian si è basato sul presupposto che chi vuole "addomesticare" i dati ha una preferenza ad utilizzare i numeri distribuiti in modo non "naturale", ossia non conformi a Benford's Law.

L'anno successivo, Varian, sebbene riconosca e definisca egli stesso BL come una legge "nebbiosa", ne identifica le potenzialità e ne ispira l'utilizzo (primo studioso a farlo) anche in ambito economico. Sempre sul medesimo assunto, lo studioso attribuisce alla Legge anche un carattere predittivo come dimostra in seguito nella descrizione del BASS Model IV (*Bay Area Simulation Study*).

- ***Raimi R.A.*[7]**

Nel proprio studio del 1976, l'autore effettua una raccolta bibliografica degli studi effettuati fino a quel momento dai propri predecessori, con lo scopo di rivedere tutte le spiegazioni alla BL proposte nei diversi settori di applicazione, rendendo chiare le ipotesi ed i risultati, con il fine ultimo di dare sostegno metodologico all'analisi.

- ***Becker P.*[8]**

Nel 1982 anche Becker sperimenta l'applicazione della BL e osserva che "*the decimal parts of failure (hazard) rates often have a logarithmic distribution*".

- ***Raimi R.A.*[9]**

L'elaborato del 1985 di Raimi pone in risalto un'idea comune nella cultura di quel periodo storico: «*la Legge di Benford è semplicemente il risultato del nostro modo di scrivere i numeri, quindi strettamente legata al nostro sistema di numerazione*».

- ***Feldstein A. e Turner P.*[10]**

Nel 1986 gli autori si specializzano nei calcoli scientifici inerenti l'assunzione di log-mantisse aritmicamente distribuite e approfondiscono lo studio della BL dal punto di vista metodologico.

- ***Schatte P.*[11]**

Nel 1988 l'autore con la pubblicazione di: "*In the course of a sufficiently long computation in floating-point arithmetic, the occurring mantissas have nearly logarithmic distribution*" fornisce

[7]Raimi, R.A. (1976). The first digit problem. American Mathematical Monthly Vol. 83 No. 7, pp. 521-538. Published by: Mathematical Association of America. Stable URL: http://www.jstor.org/stable/2319349.

[8]Becker, P. (1982) Patterns in listings of failure-rate & MTTF values and listings of other data, IEEE Transactions on Reliability R-31, 132–134.

[9]Raimi, R.A. (1985). The first digit phenomenon again. Proceedings of the American Philosophical Society, Vol. 129, No. 2, pp. 211-219.

[10]Feldstein, A. and Turner, P. (1986) Overflow, underflow, and severe loss of significance in floating-point addition and subtraction, IMA J. Numerical Analysis 6, 241–251.

[11]Schatte, P. (1988). On Benford's law to variable base. Statistics and Probability Letters 37, 391-397.

un'ampia prova che la legge della prima cifra significativa possa avere riscontro positivo anche quando applicata ai dati contabili.

- ***Thomas Jacob K.*** [12]

Nel 1989 gli studi di Thomas riprendono la ricerca effettuata da Carslaw[13] nel 1988 sui guadagni delle ditte della Nuova Zelanda, utilizzando come campione le ditte Statunitensi.

L'assunto di base che accomuna i due *Work Papers* è che nello studio della *Second-Digit*, ovvero della seconda cifra significativa, si osservano più cifre 0 e meno 9 di quelli attesi.

Il campione è strutturato considerando le ditte di COMPUSTAT[14] e i loro guadagni.

La conclusione del test "share-earning" (EPS[15]) suggerisce che l'arrotondamento in eccesso delle cifre finali è il "comportamento dominante".

Il risultato trova poi riscontro nel fatto che la quasi totalità delle aziende che dichiarano un profitto, presentano proporzioni insolitamente elevate di numeri di EPS divisibili per dieci e cinque centesimi.

Le risultanze conducono poi l'autore a cercare risposta all'ulteriore quesito: "l'arrotondamento è limitato ai casi in cui la cifra 9 è nella seconda posizione del numero?" Non necessariamente.

Lo studio conduce all'osservazione che l'arrotondamento a 0 possa verificarsi anche per le seconde cifre diverse da 9. L'arrotondamento a 0 si verifica anche quando la seconda cifra è 8, statisticamente il 5% delle volte.

La stessa osservazione è rivolta ai risultati legati alle perdite: in questo caso lo studioso nota che, al contrario di quanto avveniva per i guadagni, nella seconda cifra ci sono più 9 e meno 0 di quelli previsti. Questo è intuitivamente legato al fatto che le ditte evitano di arrotondare (e quindi cercare la cifra tonda) quando si tratta di perdite.

- ***1995 - 2005***

[12]Thomas, J.K. (1989). Unusual patterns in reported earnings. *The Accounting Review* LXIV(4), 773-787.

[13]Carslaw, C. (1988) Anomalies in Income Numbers: Evidence of Goal Oriented Behavior. *The Accounting Review* LXIII, No. 2, 321-327.

[14]Compustat is a database of financial, statistical and market information on active and inactive global companies throughout the world. The service began in 1962.

[15]EPS, acronimo di Earning per Share.

- ***Theodore P. Hill***[16,17]

Nel 1995 l'autore realizza i seguenti studi:

- Marzo: Base-Invariance Implies Benford's Law, Proceedings of the American Mathematical Society, Vol. 123, No. 3 (Mar., 1995), pp.887-895
- Aprile: The Significant-Digit Phenomenon, The American Mathematical Monthly, Vol. 102, No. 4 (Apr., 1995), pp. 322-327
- Novembre: A Statistical Derivation of the Significant-Digit Law. Statistical Science, Vol. 10, No. 4 (Nov., 1995), pp. 354-363.

Il contributo di Hill alla BL nei primi due lavori riguarda l'invarianza di scala e l'invarianza di base. Mentre nel terzo elaborato definisce con maggiore precisione le regole di BL, in particolare le caratteristiche inerenti la *Mantissa* dei numeri e la probabilità condizionata.

Secondo l'autore se congiuntamente:

- si selezionano casualmente delle distribuzioni, che godono dell'invarianza di base e di scala,
- si scelgono dei campioni casuali da ciascuna di queste distribuzioni,

allora le frequenze relative della prima cifra significativa dei numeri che si ottengono unendo i campioni casuali rispettano la BL.

In termini esemplificativi il concetto è il seguente:

1. si selezionano centomila numeri casuali tra 1 e 999.999. Si evince che la prima cifra di questi numeri si distribuisce in maniera uniforme (quindi non segue BL);
2. si suppone che il primo numero casuale sia 724.353; si sceglie quindi la seconda distribuzione tra 1 e 724.353 e si ottiene 34.121;
3. la terza distribuzione, compresa tra 1 e 34.121, genera 33.998 e così via.
4. i valori ottenuti sono abbastanza grandi per evitare che i risultati vengano "schiacciati": partendo da un insieme di numeri casuali, già al terzo passo si ha una distribuzione di probabilità delle prime cifre che segue quasi esattamente la legge di Benford.

Al termine della propria ricerca l'autore comprende che non tutti i *dataset* seguono la BL e la "Legge di Hill".

[16]Theodore Preston Hill (28 dicembre 1943) statistico statunitense.
Fonte http://www.americanscientist.org/authors/detail/theodore-hill.

[17] Ricerche e Studi dell'autore:

- Hill, T.P. (1995a). Base-invariance implies Benford's law. Proceedings of the American Mathematical Society 123, 887-895.
- Hill, T.P. (1995b). The significant-digit phenomenon. Amer. Math. Monthly 102, 322-326.
- Hill, T.P. (1995c). A statistical derivation of the significant-digit law. Statistical Science 10, 354-363.

- ***Ley E.***[18]

Nel 1996 l'autore verifica se la distribuzione delle frequenze relative della prima cifra significativa dei rendimenti giornalieri di due indici azionari americani (ovvero lo S&P per il periodo 1926-1993 e l'indice Dow Jones per il periodo 1900-1993) segue BL.

Come risultato ottiene che entrambi gli indici azionari rispettano la suddetta Legge Logaritmica.

- ***Mark j. Nigrini and l. Mittermaier***[19]

Nel 1997 gli autori conducono uno studio che introduce e descrive le "prove digitali e numeriche" che possono essere usate dai revisori in ambito di *fraud detection*, ovvero utilizzare BL per testare l'autenticità di *dataset* numerici comparando i valori reali alle frequenze digitali attese.

A tal fine, i risultati dovrebbero aiutare i revisori stessi a comprendere la veridicità dei dati analizzati.

Il *dataset* utilizzato per la ricerca appartiene ad una compagnia petrolifera quotata al NYSE. L'analisi si riferisce alle fasi di pianificazione della revisione annuale di borsa e si concentra su come determinare se i dati oggetto dell'audit contengono eccessi di cifre specifiche, combinazioni di cifre, determinati numeri o arrotondamenti. L'analisi è suddivisa per sezioni:

- la distruzione della prima cifra significativa (FIRTS DIGIT TEST)
- la distruzione della seconda cifra significativa (SECOND DIGIT TEST)
- la distruzione della combinazione delle due cifre significative (FIRST-TWO DIGIT TEST)
- la duplicazione dei dati (NUMBER DUPLICATION)
- la distribuzione della combinazione delle ultime due cifre (LAST-TWO DIGIT TEST).

Nello specifico, il *dataset* riguarda 30.084 fatture (riportanti gli importi in dollari, il beneficiario, il codice spesa, il centro di costo e l'indirizzo del beneficiario) e non contiene, perché eliminati:

- I numeri negativi, al fine di analizzare separatamente numeri positivi e negativi per evitare compensazioni,
- I numeri inferiori a $10, in quanto il i software assegna 0 come seconda cifra ai singoli *digit* interi. Ad esempio, il numero 5 viene analizzato come 5.0 con conseguente sopravvalutazione della seconda cifra 0 nei conteggi di frequenza.

[18]Ley, E. (1996). On the peculiar distribution of the U.S. Stock Indices Digits. The American Statistician Vol. 50 (No. 4), pp. 311-313. *American Statistical Association*. Stable URL: http://www.jstor.org/stable/268492.

[19]Nigrini, M.J. and L. Mittermaier (1997). The use of Benford's law as an aid in analytical procedures. *Auditing : A Journal of Practice and Theory* 16(2), 52-67. Cfr. Figure nn. 4, 5 e 6.

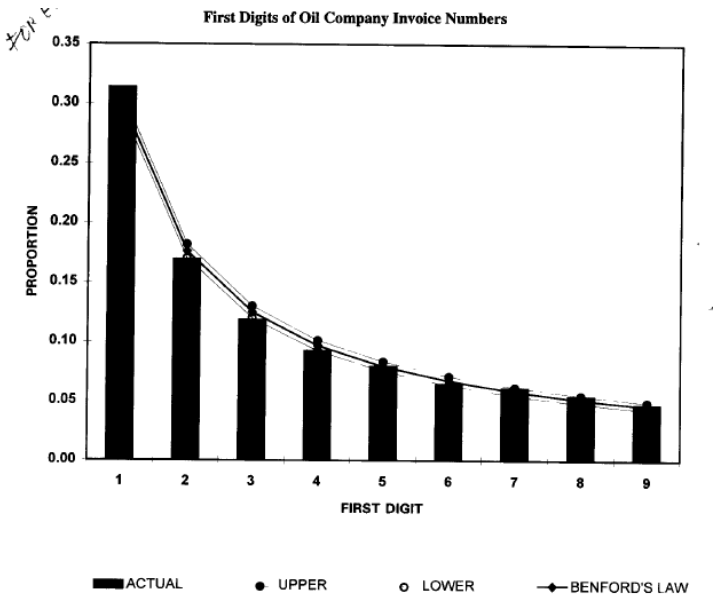

Figura 4. Distribuzione dei First Digit [20]

Il grafico (Figura 4) rappresenta il risultato della prima sezione dell'analisi "FIRST DIGIT", dove le frequenze attese della BL sono mostrate dalla linea centrale scura e le frequenze osservate sono indicate dagli istogrammi.

Le frequenze indicate dalla linea sottile sopra e sotto la linea di Benford sono i limiti superiori e inferiori di una differenza statisticamente significativa ($p < 0.01$) come misurata dallo z-test.

Lo z-test in questo ambito serve per determinare se le differenze tra le proporzioni osservate e attese sono significative (cfr. Nigrini, 1996, 80, Thomas 1989, 775). Solo le differenze per i numeri 5 e 9 non sono significative al livello 0,01.

Se il numero di osservazioni, *n*, aumenta, allora i limiti si avvicinano alle proporzioni attese. Data la bassa MAD[21] (0,44%) l'autore conclude che le prime cifre sono conformi a BL.

Nella sezione 'SECOND DIGIT' si approfondisce il Second Digit Test, che si utilizza come verifica preliminare di ragionevolezza dell'analisi.

Tuttavia, non si ritiene possa essere un modo efficiente per selezionare un campione di dati a causa delle elevate proporzioni reali.

[20]Nigrini, M.J. and L. Mittermaier (1997). The use of Benford's law as an aid in analytical procedures. *Auditing : A Journal of Practice and Theory* 16(2), 52-67.

[21]Mad, acronimo di deviazione mediana assoluta.

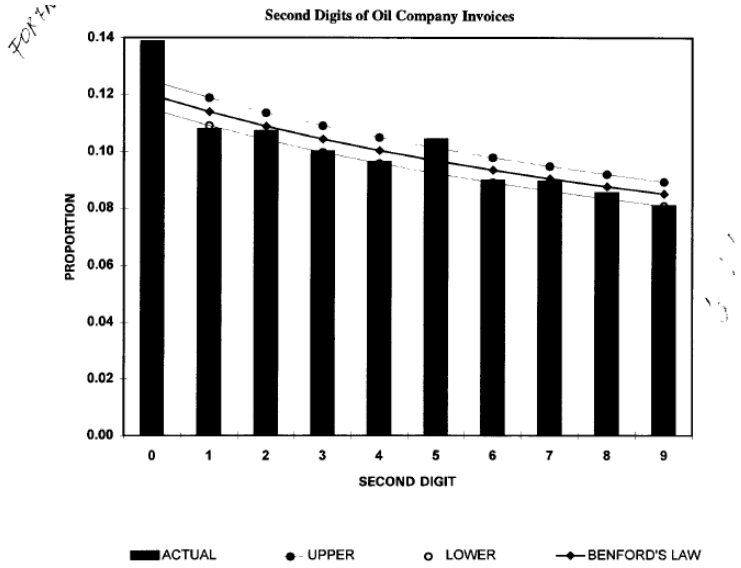

Figura 5. Distribuzione dei Second Digit

La Figura 5 mostra che la seconda cifra 0 e la seconda cifra 5 hanno frequenze effettive che superano quelle attese da BL. La deviazione più grande è dell'1,9% relativamente alla seconda cifra 0 con un MAD del 0,53%.

Dal momento che le fatture rappresentano i prezzi di beni e servizi di vendita, l'autore non si sorprende che vi sia un eccesso di *second-digit* 0 e 5, e conclude affermando che le seconde cifre sono conformi a BL e hanno un MAD del 0,53%.

Nella sezione 'FIRST-TWO DIGIT' l'autore approfondisce il First-Two Digit Test (di seguito anche "FTD") ovvero la distribuzione della combinazione delle prime due cifre significative intere delle fatture, rappresentato nel seguente grafico (Figura 6).

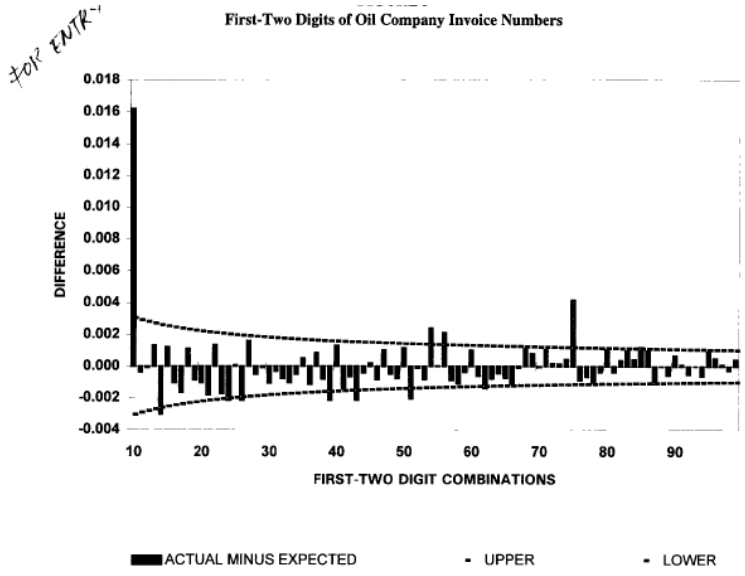

Figura 6. Distribuzione dei First-Two Digit

A titolo esemplificativo, il numero 2,204 ha una combinazione di FTD di 22. L'asse Y differisce dai grafici precedenti per il fatto che le barre mostrano la differenza tra le proporzioni reali e quelle attese.

Nell'ambito della BL, la percentuale attesa più elevata è per la FTD 10 (4.14%) mentre la percentuale minima attesa è per il FTD 99 (0,44%). Le frequenze indicate dalle linee tratteggiate sono i limiti superiori e inferiori per una significativa differenza (p <.01).

Le uniche differenze positive significative (p < 0,01) riscontrate si hanno in corrispondenza dei FTD 10, 54, 56, 75 e 85.

Nello specifico, la più grande differenza assoluta (1,6%) si ha in conformità del FTD 10. Il MAD per il set di dati è 0,11%.

L'autore ritiene che questi FTD "anomali" siano numeri sovrautilizzati e/o duplicati in un contesto di frodi, errori e inesattezze.

Nella successiva sezione 'NUMBER DUPLICATION' l'autore affronta il tema della duplicazione dei numeri ed individua alcune anomalie che segnalano la presenza d una modalità estremamente inefficiente di acquistare gli strumenti impiegati nelle operazioni di raffineria.

Nell'ultima sezione l'autore si occupa del 'LAST-TWO DIGIT' ed approfondisce il Last-Two Digit Test, ovvero la distribuzione delle ultime due cifre significative intere al fine di comprendere se vi sono arrotondamenti nei dati. In alcuni casi trova riscontro positivo e si allinea al pensiero di Thomas e Carslaw del 1988.

- ***Philip D. Drake, Mark J. Nigrini[22]***

Lo studio di Drake e Nigrini si concentra sull'individuazione delle caratteristiche che un *dataset* deve avere per essere conforme all'analisi di Benford, individuando ulteriori criteri.

Nello specifico la BL si applica:

- su grandi raccolte di dati (oltre 10.000 osservazioni) con riferimento ad un determinato periodo di tempo, come un mese, un trimestre o un anno fiscale,
- agli elenchi di numeri che descrivono fenomeni simili, come i valori di mercato o il reddito netto e alle popolazioni di città o province,
- su dati senza aggregazioni (a livello aziendale è necessario costruire il *dataset* fattura dopo fattura, senza associazioni fra centri di costo o ricavo).

[22] *Drake, P.D. and M.J. Nigrini* (2000). Computer assisted analytical procedures using Benford's law. Journal of Accounting Education 18, 127-146.

Mentre, la BL non si applica:

- agli elenchi di numeri con un valore minimo o massimo arbitrario, ovvero un *cut-off* come per esempio "città con una popolazione di oltre 50.000 persone",
- ai numeri assegnati, come i numeri di telefono o i numeri di targa dell'automobile,
- su *dataset* che comprendono numeri sia negativi sia positivi, tali valori devono essere analizzati separatamente in quanto soggetti a diversi tipi di logiche.

In questa pubblicazione gli autori, al fine di verificare se un *dataset* è conforme o meno a BL, studiano anche il test statistico del MAD per definire le soglie critiche di accettabilità dei valori da sottoporre all'analisi.

Suppongono che se la percentuale effettiva di numeri con prima cifra 1 è 0,320 (la percentuale attesa da BL è 0.301) allora la deviazione, in termini assoluti, è 0,019. Le deviazioni assolute dei 9 *digit* vengono sommati e poi divisi per 9 per ottenere il MAD.

Il MAD è applicato ai test di ragionevolezza del First Digit, Second Digit e Last Two Digit.

Per ognuno di questi test sono determinate delle griglie di conformità, riportate nelle figure sottostanti[23].

MAD: 0.000–0.004 (close conformity)
MAD: 0.004–0.008 (acceptable conformity)
MAD: 0.008–0.012 (marginally acceptable conformity)
MAD: greater than 0.012 (nonconformity)

Figura 7. Conformità MAD - *First Digit*. La figura mostra i livelli di conformità del MAD: con deviazione media assoluta superiore allo 0,012 il *dataset* viene dichiarato non conforme al test della prima cifra.

MAD: 0.000–0.008 (close conformity)
MAD: 0.008–0.012 (acceptable conformity)
MAD: 0.012–0.016 (marginally acceptable conformity)
MAD: greater than 0.016 (nonconformity)

Figura 8. Conformità MAD - *Second Digit*. La figura mostra i livelli di conformità del MAD: con deviazione media assoluta superiore allo 0,016 il *dataset* viene dichiarato non conforme al test della seconda cifra.

MAD: 0.0000–0.0006 (close conformity)
MAD: 0.0006–0.0012 (acceptable conformity)
MAD: 0.0012–0.0018 (marginally acceptable conformity)
MAD: greater than 0.018 (nonconformity)

Figura 9. Conformità MAD – *First-Two Digit.* La figura mostra i livelli di conformità del MAD: con deviazione media assoluta superiore allo 0,018 il *dataset* viene dichiarato non conforme al test della prime due cifre significative.

[23]Tratte da *Drake, P.D. and M.J. Nigrini* (2000). Computer assisted analytical procedures using Benford's law. *Journal of Accounting Education* 18, 127-146.

- ***Bassam Hasan, Ph.D.*** [24]

Nel 2002 l'autore segue il filone di studi ed applicazioni della BL come metodo di previsione e propone un esempio pratico di riscontro di BL nella realtà: la crescita della popolazione.

Egli parte dalle seguenti ipotesi:

- il numero della popolazione in una contea degli Stati Uniti è di 10.000 persone,
- la popolazione cresce ad un tasso annuo del 2%.

Il risultato che ottiene rileva che sono necessari circa 36 anni ad una sola contea per poter raddoppiare la propria popolazione a 20.000 abitanti.

L'autore si spiega osservando che tale numero inizierà con un 1 per circa 36 anni ed il successivo cambiamento della cifra iniziale si verifica quando il numero della popolazione raggiunge i 30.000 abitanti. Cosi gradualmente per il successivo cambiamento da 2 a 3, per il quale servono circa 20 anni. Allo stesso modo occorre attendere circa 14 anni per il cambiamento da 3 a 4, e 6 anni affinché la prima cifra passi da 8 a 9.

L'autore conclude affermando che

- la variazione della cifra iniziale che ha richiesto maggiore tempo è quella compresa tra 1 e 2,
- la probabilità che la contea abbia un numero della popolazione che inizia con 1 è circa nove volte superiore alla probabilità che il numero inizi con 9.

- ***Zhipeng L., Lin C. and Huajia W.***[25]

Gli autori si occupano di presentare tre test da adottare per esaminare l'idoneità dei *dataset* alla distribuzione di Benford: Test Chi-Quadrato, '*total variation distance*', e il '*maximum deviations*'.

- ***Sehity T., Hoelzl E., Kirchler E.*** [26]

Nel 2005 gli autori utilizzano la BL per studiare i prezzi al consumo prima e dopo l'introduzione della moneta "euro".

Si riporta un estratto della pubblicazione degli autori: «*Retail managers use psychological pricing to make the prices of goods appear to be just below a round number. The euro introduction in 2002, with its various exchange rates, distorted existing nominal price patterns while at the same time retaining real prices. We studied consumer prices before and after the introduction of the euro by using Benford's Law as a benchmark for price adjustments. Results indicate the usefulness of this*

[24]Tratto da "Assessing Data Authenticity with Benford's Law" Bassam Hasan, Ph.D. (2002).

[25]Zhipeng L., Lin C. and Huajia W. (2004). On Wilson's theorem and Polignac conjecture. http://arxiv.org/abs/math.NT/0408018.

[26]Sehity, T., Hoelzl, E., Kirchler, E. (2005). Price developments after a nominal shock: Benford's Law and psychological pricing after the euro introduction. International Journal of Research in Marketing 22(4), 471-480. ISSN:0167-8116.

benchmark for detecting irregularities in prices, and a clear trend towards psychological pricing after the nominal shock of the euro introduction. In addition, the tendency towards psychological prices results in different inflation rates in dependence of the price pattern.»

- ***2006 - 2015***
  - ***Saville A.D.*** [27]

Nel 2006 l'autore utilizza i dati tratti da società quotate alla Borsa di Johannesburg per verificare l'ipotesi che la BL possa essere utilizzata per identificare segnalazioni false o fraudolente dei dati contabili. I risultati riscontrano positivamente l'ipotesi.

Di conseguenza, le considerazioni sono di particolare importanza in diversi campi applicativi e per differenti figure professionali (revisori, azionisti, analisti finanziari, gestori di investimenti, investitori).

  - ***Miller S.J and Nigrini M.J.*** [28]

Nel 2007 Nigrini e Miller approfondiscono lo studio di BL anche al fine di poterne far impiego per "salvaguardare il pianeta" e nel misurare gli effetti delle azioni dell'uomo sulla natura.

  - ***Douglas N. Hales, Satya S. Chakravorty, V. Sridharan*** [29]

Nel 2008 gli autori utilizzano la BL nell'ambito della Supply Chain e osservano che la stessa può essere utilizzata anche per l'individuazione di fornitori "opportunisti".

  - ***Wang J., Cha B., Cho S. and Kuo C.*** [30]

Nel 2009 gli autori presentano lo studio della BL nell'ambito delle immagini digitali e notano che esiste una relazione tra queste, essendo i dati conformi all'analisi di Benford.

  - ***Hill T.P. and Berger A.*** [31,32,33]

[27]Saville, A. (2006). Using Benford's Law to detect data error and fraud: An examination of companies listed on the Johannesburg Stock Exchange. *South African Journal of Economic and Management Sciences*, Vol. 9(3), 341-354. Web. Http://www.scribd.com/doc/47789223/Saville-Using-282006-29.

[28]Nigrini, M., & Miller, S. (2007). Benford's Law applied to hydrology data--results and relevance to other geophysical data. *Mathematical Geology*, 39(5), 469-490.

[29]Hales, D.N., Sridharan, V., Radhakrishnan, A., Chakravorty, S.S. and Sihad, S.M. (2008). Testing the accuracy of employee-reported data: An inexpensive alternative approach to traditional methods. *European Journal of Operational Research* 189(3), 583-593.

[30]Wang, J., Cha, B., Cho, S. and Jay Kuo C. (2009). Understanding Benford's Law and its vulnerability in image forensics. Signal and Image Processing Institute and Ming Hsieh Department of Electrical Engineering University of Southern California, Los Angeles, CA 90089-2564.

[31]Berger, A., and Hill, T.P.(2011). A basic theory of Benford's Law . *Probability Surveys* 8, 1-126.

[32]Berger, A., and Hill, T.P.(2011). Benford's Law Strikes Back: No Simple Explanation in Sight for Mathematical Gem. The Mathematical Intelligencer 33(1), 85-91. DOI:10.1007/ s00283-010-9182-3.

[33]Berger, A. and Hill, T.P., Kaynar, B. and Ridder, A. (2011). Finite-state Markov Chains Obey Benford's Law. SIAM *Journal of Matrix Analysis and Applications* 32(3), 665-684.

Nel 2011 gli autori cercano di fornire ulteriori spiegazioni alla BL rispetto ai precedenti studi[34] ed esaminano gli "errori" più ricorrenti nell'applicazione della stessa, i c.d. ''back-of-the-envelope''[35] .
Gli autori ritengono che, nonostante le principali caratteristiche di BL siano state individuate e verificate, non vi sia alcun approccio unificato che spiega al contempo la sua comparsa nei sistemi dinamici, nella teoria dei numeri, nelle statistiche e nei dati del mondo reale, con la conseguenza che l'applicazione di BL, soprattutto nei dati reali, rimane misteriosa.

- ***Zaharis A., Martini A., Tryfonas T., Illioudis C. and Pangalos G.[36]***

Nel 2011 gli autori presentano un nuovo metodo di steg-analisi delle immagini JPEG ed applicano i principi di BL per individuare anomalie nella ricostruzione delle immagini. L'approccio è guidato dalla necessità di identificare rapidamente e precisamente le anomalie in un immagine, ovvero in un insieme di file di diversi formati, dove non vi è conoscenza dell'algoritmo stenografico utilizzato. Gli autori trovano riscontro positivo in quanto la BL si presta anche a questo tipo di utilizzo.

- ***Zgela, M. and Dobša, J.[37]***

Gli autori esaminano le modalità di applicazione di BL in un'indagine sul reddito netto (utile o perdita) delle prime 500 aziende dell'Europa centrale e orientale nel triennio 2007-2009. L'obiettivo di individuare anomalie nei dataset viene raggiunto in quanto riscontrano discrepanze per alcune prime cifre, in particolare in corrispondenza della cifra 5 e della cifra 9.
Gli autori approfondiscono lo studio e ritengono che, indipendentemente da questo, i risultati del test chi-quadrato forniscono prova che tutti i sottoinsiemi di dati individuati sono conformi con la BL.

[34] Precedenti studi dell'autore Hill, T.P.:
- Hill, TP (1996). A note on distributions of true versus fabricated data. Perceptual and Motor Skills 83, 776-778 Part 1. ISSN:0031-5215.
- Hill, T.P. (1997). Benford's law. Encyclopedia of Mathematics Supplement, vol. 1, 102.
- Hill, T.P. (1998). The first digit phenomenon. (A century-old observation about an unexpected pattern in manynumerical tables applies to the stock market, census statistics and accounting data), The American Scientist, Vol. 86, No. 4 (JULY-AUGUST 1998), pp. 358-363. Published by: Sigma Xi, The Scientific Research Society. Stable URL: http://www.jstor.org/stable/27857060.
- Hill, T.P. (1999c). The difficulty of faking data. Chance 26, 8-13.
- Hill, T.P. and K. Schürger (2005). Regularity of digits and significant digits of random variables. Journal of Stochastic Processes and their Applications 115, 1723-1743.

[35]Traduzione di ''back-of-the-envelope'': "calcolo fatto grossolanamente".

[36]Zaharis A., Martini A., Tryfonas T., Illioudis C. and Pangalos G.,(2011). Reconstructive Steganalysis by Source Bytes Lead Digit Distribution Examination. 1University of Thessaly, Greece SIEMENS SA, Greece, University of Bristol, UK, ATEI of Thessaloniki, Greece, Aristotle University of Thessaloniki, Greece.

[37]Žgela, M. and Dobša, J. (2011). Analysis of Top 500 Central and East European Companies Net Income Using Benford's Law. *Journal of Information and Organizational Sciences*, Vol.35, No.2, Prosinac 2011, pp. 215-228.

- ***Joenssen D.W.***[38,39]

Nel 2013 l'autore propone una variazione nell'applicazione di BL: propone di analizzare non solo la distribuzione delle prime cifre, ma la distribuzione congiunta delle prime due cifre dei dati al fine di ottenere un effetto rafforzativo dell'esito dell'analisi. Infine ne confronta i risultati ottenuti. Nella propria ricerca l'autore propone inoltre quattro test di "bontà di adattamento" dei *dataset* alla BL per verificarne l'applicabilità. I test statistici comparati sono:

- χ^2 di Pearson (Pearson, 1900),
- D di Kolmogorov-Smirnov (Kolmogorov, 1933),
- la modifica di Freedman di Watson U^2_n per distribuzioni discrete (Freedman, 1981)
- la correlazione statistica J^2_P, un tipo di test di Shapiro-Francia (Shapiro e Francia, 1972).

La conclusione a cui perviene l'autore è che i *dataset* analizzati sono conformi alla BL.

6. Conclusioni

L'analisi di Benford's Law ha ricevuto attenzione da parte di studiosi e ricercatori appartenenti a differenti settori: dalla biologia, alla chimica, alla fisica, all'informatica, all'economia e alla politica. Ciascuno di essi ha contribuito ad approfondire la conoscenza di questa legge logaritmica, che è capace di individuare anomalie all'interno di un'ampia raccolta di dati e di consentire la rappresentazione grafica dei risultati in modo intuitivo tramite la curva delle frequenze delle cifre indagate.

Tuttavia, come riscontrato da alcuni ricercatori, non tutti i *dataset* rispettano la Benford's Law ed in quanto tali non sono sottoponibili all'analisi.

Il *dataset* per essere conforme deve infatti avere determinate caratteristiche e rispettare determinati parametri, diversamente potrebbe condurre a conclusioni non corrette. Per questo motivo diviene necessario verificare tale conformità prima di sottoporlo alla Benford's Law. In questo modo si migliora anche il grado di affidabilità dei risultati dell'analisi.

[38]Joenssen D.W. (2013). BenfordTests: Statistical Tests for Evaluating Conformity to Benford's Law. R package version 1.1.1.

[39]Joenssen, DW (2013). Two digit testing for Benford's Law. *Proceedings of the ISI World Statistics Congress*, 59th Session in Hong Kong.

Bibliografia

Bassam Hasan, Ph.D. (2002). “Assessing Data Authenticity with Benford's Law”.

Becker, P. (1982) Patterns in listings of failure-rate & MTTF values and listings of other data, IEEE *Transactions on Reliability* R-31, 132–134.

Benford Frank, (1938), The law of anomalous numbers, Physicist, Research Laboratory, General Electric Company, Schenectadt, New York, *American Philosophical Society.*

Berger, A., and Hill, T.P.(2011). A basic theory of Benford's Law . *Probability Surveys* 8, 1-126.

Berger, A., and Hill, T.P.(2011). Benford's Law Strikes Back: No Simple Explanation in Sight for Mathematical Gem.

Berger, A. and Hill, T.P., Kaynar, B. and Ridder, A. (2011). Finite-state Markov Chains Obey Benford's Law. SIAM *Journal of Matrix Analysis and Applications* 32(3), 665-684.

Carslaw, C. (1988) Anomalies in Income Numbers: Evidence of Goal Oriented Behavior. *The Accounting Review* LXIII, No. 2, 321-327.

Drake, P.D. and M.J. Nigrini (2000). Computer assisted analytical procedures using Benford’s law. *Journal of Accounting Education* 18, 127-146

Feldstein, A. and Turner, P. (1986) Overflow, underflow, and severe loss of significance in floating-point addition and subtraction, IMA J. *Numerical Analysis* 6, 241–251.

Hales, D.N., Sridharan, V., Radhakrishnan, A., Chakravorty, S.S. and Sihad, S.M. (2008). Testing the accuracy of employee-reported data: An inexpensive alternative approach to traditional methods. *European Journal of Operational Research* 189(3), 583-593

Hassan, B. (2002). Assessing data authencity with Benford’s law. Information Systems Control Journal 6.

Hill, T.P. (1995a). Base-invariance implies Benford’s law. *Proceedings of the American Mathematical Society* 123, 887-895.

Hill, T.P. (1995b). The significant-digit phenomenon. Amer. Math. Monthly 102, 322-326.

Hill, T.P. (1995c). A statistical derivation of the significant-digit law. *Statistical Science* 10, 354-363.

Hill, TP (1996). A note on distributions of true versus fabricated data. *Perceptual and Motor Skills* 83, 776-778 Part 1. ISSN:0031-5215.

Hill, T.P. (1997). Benford’s law. *Encyclopedia of Mathematics Supplement*, vol. 1, 102.

Hill, T.P. (1998). The first digit phenomenon. (A century-old observation about an unexpected pattern in manynumerical tables applies to the stock market, census statistics and accounting data),

The American Scientist, Vol. 86, No. 4 (JULY-AUGUST 1998), pp. 358-363. Published by: Sigma Xi, The Scientific Research Society. Stable URL: http://www.jstor.org/stable/27857060.

Hill, T.P. (1999c). The difficulty of faking data. Chance 26, 8-13.

Hill, T.P. and K. Schürger (2005). Regularity of digits and significant digits of random variables. *Journal of Stochastic Processes and their Applications* 115, 1723-1743.

Joenssen D.W. (2013). BenfordTests: Statistical Tests for Evaluating Conformity to Benford's Law. R package version 1.1.1.

Joenssen, DW (2013). Two digit testing for Benford's Law. *Proceedings of the ISI World Statistics Congress*, 59th Session in Hong Kong.

Knuth, D. (1969). The Art of Computer Programming, vol. 2, 219-229. (2nd ed.). Addison-Wesley, Reading, MA, 239-249. (3rd ed.) (1998), 254-262.

Ley, E. (Nov., 1996). On the peculiar distribution of the U.S. Stock Indices Digits. *The American Statistician* Vol. 50 (No. 4), pp. 311-313. American Statistical Association. Stable URL: http://www.jstor.org/stable/268492.

Newcombe Simon, Note on the Frequency of Use of the Different Digits in Natural Numbers, "*American Journal of Mathematics"*, Vol. 4, No. 1. (1881), pp. 39-40.

Nigrini, M., & Miller, S. (2007). Benford's Law applied to hydrology data--results and relevance to other geophysical data. *Mathematical Geology*, 39(5), 469-490.

Nigrini, M.J. and L. Mittermaier (1997). The use of Benford's law as an aid in analytical procedures. *Auditing : A Journal of Practice and Theory* 16(2), 52-67

Pinkham, R.S. (1961). On the distribution of first significant digits. *Annals of Mathematical Statistics* 32, 1223-1230.

Raimi, R.A. (1976). The first digit problem. American Mathematical Monthly Vol. 83 No. 7, pp. 521-538. Published by: Mathematical Association of America. Stable URL: http://www.jstor.org/stable/2319349.

Raimi, R.A. (1985). The first digit phenomenon again. *Proceedings of the American Philosophical Society*, Vol. 129, No. 2, pp. 211-219.

Saville, A. (2006). Using Benford's Law to detect data error and fraud: An examination of companies listed on the Johannesburg Stock Exchange. *South African Journal of Economic and Management Sciences*, Vol. 9(3), 341-354. Web. Http://www.scribd.com/doc/47789223/Saville-Using-282006-29.

Schatte, P. (1988). On Benford's law to variable base. Statistics and Probability Letters 37, 391-397.

Thomas, J.K. (1989). Unusual patterns in reported earnings. *The Accounting Review* LXIV(4), 773-787.

Sehity, T., Hoelzl, E., Kirchler, E. (2005). Price developments after a nominal shock: Benford's Law and psychological pricing after the euro introduction. International Journal of Research in Marketing 22(4), 471-480. ISSN:0167-8116.

Varian, Hal R. (1972). "Benford's Law (Letters to the Editor)". *The American Statistician* 26 (3): 65. doi:10.1080/00031305.1972.10478934.

Wang, J., Cha, B., Cho, S. and Jay Kuo C. (2009). Understanding Benford's Law and its vulnerability in image forensics. Signal and Image Processing Institute and Ming Hsieh Department of Electrical Engineering University of Southern California, Los Angeles, CA 90089-2564.

Zaharis A., Martini A., Tryfonas T., Illioudis C. and Pangalos G.,(2011). Reconstructive Steganalysis by Source Bytes Lead Digit Distribution Examination. 1University of Thessaly, Greece SIEMENS SA, Greece, University of Bristol, UK, ATEI of Thessaloniki, Greece, Aristotle University of Thessaloniki, Greece.

Žgela, M. and Dobša, J. (2011). Analysis of Top 500 Central and East European Companies Net Income Using Benford's Law. *Journal of Information and Organizational Sciences*, Vol.35, No.2, Prosinac 2011, pp. 215-228.

Zhipeng L., Lin C. and Huajia W. (2004). On Wilson’s theorem and Polignac conjecture. http://arxiv.org/abs/math.NT/0408018.