



IVASS
ISTITUTO PER LA VIGILANZA
SULLE ASSICURAZIONI



Quaderno n. 19

Test di Benford sulla qualità dei dati

Riccardo Cesari



Ottobre 2021

(decreto legge 6 luglio 2012 n. 95 convertito con legge 7 agosto 2012 n. 135)

La serie Quaderni intende promuovere la diffusione di studi e contributi originali sui temi assicurativi al fine di suscitare commenti critici e suggerimenti.

Le opinioni espresse nei lavori sono attribuibili ai soli autori e non impegnano in alcun modo la responsabilità delle Istituzioni di appartenenza.

via del Quirinale 21 - 00187
ROMA telefono +39 06 42133.1

Tutti i diritti riservati.

È consentita la riproduzione a fini didattici e non commerciali,
a condizione che venga citata la fonte

La serie è disponibile online nel sito www.ivass.it

ISSN 2421-4671 (online)

Test di Benford sulla qualità dei dati

Da tempo, la c.d. Legge di Benford è stata usata come un indicatore di dubbia qualità dei dati (v. Varian, 1972 e l'Appendice per alcuni cenni storici).

Infatti, per quanto possa sembrare strano, Frank Benford (1938), riscontrò, su vari ed eterogenei gruppi di dati, una distribuzione non uniforme della prima cifra dei numeri.

Invece di trovare, come si potrebbe supporre, una frequenza costante di $1/9=11.1\%$, la prima cifra di vari fenomeni quantitativi, naturali e non (dalla superficie delle regioni al peso di composti chimici agli indirizzi di un elenco telefonico) risultava avere una distribuzione che è molto ben approssimabile con $\log_{10}(1+p)-\log_{10}(p)$ (Tav. 1).

Tab. 1 - Distribuzione di Benford della prima cifra di un numero

p	1	2	3	4	5	6	7	8	9
Freq %	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6
Cumulata	30.1	47.7	60.2	69.9	77.8	84.5	90.3	95.4	100

Il suo uso a fini “antifrode” deriva proprio da questa grande pervasività. Infatti, qualora, in un certo fenomeno, si osservasse una distribuzione molto diversa da quella di Benford, $F_0(j)$, si potrebbe ricavare un segnale di “alert” per la possibile presenza di dati manipolati.

Il test è di facile calcolo.

Ad esempio per un'impresa si possono prendere le n segnalazioni di c/economico, vedere la frequenza della prima cifra dei vari items contabili, calcolare le frequenze cumulate $F(j)$ e applicare il test di Kolmogorov-Smirnov di somiglianza tra la distribuzione ottenuta e quella “teorica” di Tab. 1.

$$D_n = \max_{j=1,\dots,9} |F(j) - F_0(j)|$$

Un valore del test superiore alla soglia critica di Tab. 2, per dato livello di confidenza α (es. 1%) e data numerosità n delle osservazioni, indica una differenza significativa tra la distribuzione osservata e quella “teorica”.

Ad esempio, allo 0.1% di significatività risultano “anomali” (nel senso suddetto) i dati del c/economico 2019 di A105S Sara e A475S RBM.

All'1% di significatività ¹ si aggiungono A077S Zurich, A127S UCA, A247S Genertel e A478S Amtust.

Tab. 2 - Valori critici del test di Kolmogorov-Smirnov in funzione della numerosità campionaria (n) e del livello di confidenza (α) (*)

$n \backslash \alpha$	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907
10	0.58042	0.48895	0.45662	0.40925	0.36866	0.34250	0.32257
11	0.55588	0.46770	0.43670	0.39122	0.35242	0.32734	0.30826
12	0.53422	0.44905	0.41918	0.37543	0.33815	0.31408	0.29573
13	0.51490	0.43246	0.40362	0.36143	0.32548	0.30233	0.28466
14	0.49753	0.41760	0.38970	0.34890	0.31417	0.29181	0.27477
15	0.48182	0.40420	0.37713	0.33760	0.30397	0.28233	0.26585
16	0.46750	0.39200	0.36571	0.32733	0.29471	0.27372	0.25774
17	0.45440	0.38085	0.35528	0.31796	0.28627	0.26587	0.25035
18	0.44234	0.37063	0.34569	0.30936	0.27851	0.25867	0.24356
19	0.43119	0.36116	0.33685	0.30142	0.27135	0.25202	0.23731
20	0.42085	0.35240	0.32866	0.29407	0.26473	0.24587	0.23152
25	0.37843	0.31656	0.30349	0.26404	0.23767	0.22074	0.20786
30	0.34672	0.28988	0.27704	0.24170	0.21756	0.20207	0.19029
35	0.32187	0.26898	0.25649	0.22424	0.20184	0.18748	0.17655
40	0.30169	0.25188	0.23993	0.21017	0.18939	0.17610	0.16601
45	0.28482	0.23780	0.22621	0.19842	0.17881	0.16626	0.15673
50	0.27051	0.22585	0.21460	0.18845	0.16982	0.15790	0.14886
OVER 50	1.94947	1.62762	1.51743	1.35810	1.22385	1.13795	1.07275
	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$

(*) Per $n > 50$ il valore critico si ottiene dividendo il numero in ultima riga per $\sqrt{n}$

¹ Si noti che all'1% di significatività, su 101 imprese ben 98 non si conformano alla legge uniforme. Risultano conformi (meglio: statisticamente non difforni) solo Sara, UCA e A383S Credemvita.

Appendice: storia del problema

Il primo a notare la non uniformità della distribuzione della prima cifra significativa di un numero sembra essere stato Simon Newcomb (1881), a partire dalla semplice osservazione che le tavole dei logaritmi risultavano più consumate nelle prime pagine che nelle ultime.

Egli trovò, euristicamente, che la prima cifra (tra 1 e 9) ha la distribuzione cumulata di probabilità $\log_{10}(1+p)$ e la distribuzione di frequenza $\log_{10}(1+p)-\log_{10}(p)$ e ne calcolò i valori. Tuttavia, questa formalizzazione analitica si deve a Benford (1938) che chiamò la distribuzione col suo nome.

Newcomb trovò anche la distribuzione della seconda cifra nonché la tendenza all'uniformità al crescere della posizione (Tav. A1)

Tav. A1

Frequenza dei numeri in base alla posizione secondo la legge di Newcomb-Benford

Numero										
Posizione	0	1	2	3	4	5	6	7	8	9
1	-	30.10%	17.61%	12.49%	9.69%	7.92%	6.69%	5.80%	5.12%	4.58%
2	11.97%	11.39%	10.88%	10.43%	10.03%	9.67%	9.34%	9.04%	8.76%	8.49%
3	10.18%	10.14%	10.10%	10.06%	10.02%	9.98%	9.94%	9.90%	9.86%	9.82%
4	10.02%	10.01%	10.01%	10.01%	10.00%	10.00%	9.99%	9.99%	9.99%	9.98%

Una semplice spiegazione del fenomeno (Fewster, 2009) si può ricavare dalla rappresentazione dei numeri in notazione scientifica in base 10: $X(v,n) = v \times 10^n$ dove $1 \leq v < 10$. Chiaramente, la prima cifra (significativa) di X è la prima cifra di v e, in particolare, è 1 se $1 \leq v < 2$, è 2 se $2 \leq v < 3$ etc.

Prendendo il log in base 10 si ha: $\log_{10}(X) = \log_{10}(v) + n$ e $1 \leq v < 2$ se $0 \leq \log_{10}(v) < \log_{10}(2) = 0.301$;

$2 \leq v < 3$ se $0.301 \leq \log_{10}(v) < \log_{10}(3) = 0.4771$ etc. Pertanto la probabilità di cifra 1 è 0.301, di cifra 2 è $0.4771 - 0.3010 = 0.1761$ etc.

Da notare che lo stesso Newcomb (1881) suggeriva ("curious to remark") l'uso di questa legge per valutare la natura di una "large collection of independent numerical results".

Bibliografia

Benford, F. A. (1938), The law of anomalous numbers, *Proceedings of the American Philosophical Society*, 78, 551-572

Fewster, R. M. (2009), A simple explanation of the Benford's law, *The American Statistician*, 63, 1, 26-32

Newcomb, S. (1881), Note on the frequency of the use of the different digits in natural numbers, *American Journal of Mathematics*, 4, 1, 39-40

Varian, H. (1972), Benford's Law, *The American Statistician*, June, 65-66

QUADERNI PUBBLICATI

- N. 1. [Il Ramo r.c. auto: raffronto tra l'Italia e alcuni paesi della UE su premi, sinistri e sistemi risarcitori del danno alla persona](#), di Lino Matarazzo (ottobre 2014).
- N. 2. [La riforma della CARD: costi dei sinistri e incentivi all'efficienza nel risarcimento diretto r.c. auto](#), di Riccardo Cesari, Marina Mieli e Arturo Valerio (maggio 2015).
- N. 3. [Dal chain ladder al modello di Merz e Wüthrich: derivazione completa del modello di volatilità della riserva sinistri in orizzonte annuale](#), di Stefano Cavastracci (giugno 2015).
- N. 4. [Redditività e ciclo del settore assicurativo italiano prima e durante la crisi](#), di Fabio Farabullini (novembre 2015).
- N. 5. [Seminari per i dieci anni del codice delle assicurazioni private](#), di Riccardo Cesari (Prefatore), Enrico Galanti (Coordinatore), Sandro Amorosino, Roberto Caponigro, Stefania Ceci, Luigi Farenga, Antonio Longo, Francesco Mauro, Gustavo Olivieri, Andrea Pezzoli, Salvatore Providenti, Umberto Santosuosso, Dario Zamboni (maggio 2016).
- N. 6. [Il diritto nella società contemporanea](#), di Ottavio De Bertolis (novembre 2016).
- N. 7. [Duration, convexity and the optimal management of bond portfolios for insurance companies](#), di Riccardo Cesari e Vieri Mosco (febbraio 2017).
- N. 8. [Il nuovo Regolamento IVASS sull'accesso agli atti - La distribuzione Assicurativa - Il gruppo dopo Solvency II](#), di E. Galanti, M. Binda, M. L. Cavina, M. Fodale, N. Gentile, R. Giay, P. Marano, P. Mariano, S. Marzucchi, A. Police, A. Serino, V. Troiano (aprile 2017).
- N. 9. [Modello overdispersed Poisson: formula chiusa per la stima GLM della volatilità one year della riserva sinistri](#), di Stefano Cavastracci e Agostino Tripodi (giugno 2017).
- N. 10. [No news is good news: moral hazard in oligopolistic insurance markets](#), di Marco Cosconati (aprile 2018).
- N. 11. [Riforma del sistema europeo delle Autorità di controllo. Governance imprese assicurative. Gestione crisi nel settore bancario e assicurativo](#), di S. Butera, F. Buzzichelli, R. Cercone, A. Corinti, S. De Polis, C. Di Noia, S. Fortunato, E. Galanti, F. Montemaggiori, M. Morvillo, M. O. Perassi, P. Rosatone, V. Santoro, S. Scarcello, E. Serata, M. Siri (maggio 2018).

- N. 12. [Valore aggiunto e profittabilità delle compagnie italiane: cosa conta realmente?](#), di Leandro D'Aurizio (marzo 2019)
- N. 13. [Calamità naturali e coperture assicurative: valutazione dei rischi e policy options per il caso italiano](#), di Riccardo Cesari e Leandro D'Aurizio (luglio 2019)
- N. 14. [Defiscalizzare la r.c. auto: come e perché](#), di Riccardo Cesari e Antonio R. De Pascalis (ottobre 2019)
- N. 15. [Two simple models of insurance fraud](#), di: Riccardo Cesari (gennaio 2021)
- N. 16. [La governance dell'Artificial Intelligence nel settore assicurativo tra principi etici, responsabilità del board e cultura aziendale](#), di Diana Capone (febbraio 2021)
- N. 17. [The effect of uncertainty on the car insurance market: evidence from the COVID-19 shock](#), di Marco COSCONATI e Viviana MEDORI (agosto 2021)
- N. 18. [Le relazioni tra banche e assicurazioni in Italia](#), di: Federico Apicella, Leandro D'Aurizio, Raffaele Gallo, Giovanni Guazzarotti (settembre 2021)

