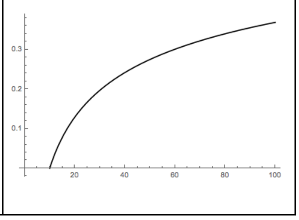
ISSN: 2037-6367

MatematicaMente

Pubblicazione mensile della sezione veronese della MATHESIS – Società Italiana di Scienze Matematiche e Fisiche – Fondata nel 1895 – Autorizzazione del Tribunale di Verona n. 1360 del 15 – 03 – 1999 – I diritti d'autore sono riservati. Direttore: Luciano Corso – Redazione: Alberto Burato, Fabrizio Giugni, Michele Picotti, Sisto Baldo – Via IV Novembre, 11/b – 37126 Verona – tel e fax (045) 8344785 – 338 6416432 – e-mail: lcorso@iol.it – info@mathesisverona.it – Stampa in proprio – Numero 230 – Pubblicato il 03 – 11 – 2017



Tributo a Cantor

Carlo Toffalori [*]

[*Segue dal numero 229*]

Quella storicamente più rilevante è comunque la seconda, che esclude la numerabilità dei reali trascendenti. Per raggiungere questo obiettivo, Cantor considera

- da un lato una qualsiasi successione di reali $r_0, r_1, r_2, \dots, r_n, \dots$ (per esempio quella dei reali algebrici),
- dall'altro un qualunque intervallo di $\mathbb{R}$, diciamo di estremi $a < b$,

e dimostra come questo intervallo ospiti elementi estranei alla successione, in particolare, nel caso dei reali algebrici, elementi trascendenti. Ma il discorso, come si preannunciava, è assai più generale e stabilisce che nessun intervallo e neppure l'intero $\mathbb{R}$ sono numerabili, perché non c'è modo di esaurirli con una successione come quella degli r_n. L'infinità di $\mathbb{R}$ è perciò distinta da quella di $\mathbb{N}$.
Quanto poi ai numeri trascendenti, aggiungiamo che, se anch'essi si potessero arrangiare in una successione numerabile come gli algebrici,

- $t_0, t_1, t_2, \dots, t_n, \dots$ gli uni e
- $s_0, s_1, s_2, \dots, s_n, \dots$ gli altri,

allora la successione che li alterna $t_0, s_0, t_1, s_1, t_2, s_2, \dots, t_n, s_n, \dots$ conterrebbe tutti i reali e dunque ne riempirebbe ogni intervallo, in contraddizione col risultato appena enunciato. Dunque pure i reali trascendenti ammettono un'infinità diversa dal numerabile – coincidente, anzi, col continuo.
Arriviamo finalmente ai dettagli della dimostrazione di Cantor, che parte appunto dai due estremi $a < b$ e si configura come una sorta di gioco tra due contendenti immaginari I e II:

- I comincia la partita indicando tra gli r_n il primo elemento, chiamiamolo a_0, strettamente compreso tra a e b, dunque r_0 se $a < r_0 < b$, e altrimenti r_1, o r_2, o altri successivi – notiamo che, se I fallisce nel suo scopo, allora non c'è addirittura nessun reale della successione all'interno dell'intervallo di estremi a e b, e quindi la tesi è dimostrata;
- II risponde indicando, se c'è, il primo degli r_n, diciamo b_0, tale che $a_0 < b_0 < b$;
- I gli contrappone il primo tra gli r_n rimasti (chiamiamolo a_1) tale che $a_0 < a_1 < b_0$;
- II oppone il primo degli r_n residui, b_1, per cui $a_1 < b_1 < b_0$,

e così via. In questo modo l'intervallo di ricerca si restringe successivamente da $]a, b[$ a $]a_0, b[$, $]a_0, b_0[$, $]a_1, b_0[$, $]a_1, b_1[$ eccetera. Ogni volta il giocatore che muove, I o II, cerca se possibile il primo r_n all'interno del nuovo intervallo. Si noti allora che gli a_n formano una sequenza crescente superiormente limitata da b e, similmente, i b_n formano una sequenza decrescente inferiormente limitata da a, si ha anzi che $a < a_0 < a_1 < \cdots < b_1 < b_0 < b$. Inoltre r_0, cioè il primo elemento della successione originaria,

- o sta fuori di $]a, b[$, nel qual caso non può essere selezionato né da I né da II in nessuno degli intervalli sempre più stretti che si vengono a generare,
- oppure sta all'interno di $]a, b[$ e allora viene scelto come a_0 e resta di conseguenza escluso dai passaggi successivi.

Allo stesso modo nessun r_n può essere coinvolto nei turni che seguono la $(n+1)$-ma mossa. In altre parole ogni r_n resta fuori dall'intervallo di posto $n+2$ della costruzione e quindi da tutti i successivi.
A questo punto i casi sono due:

- o il gioco si interrompe dopo un numero finito di passi, come dire che nell'ultimo intervallo considerato non c'è più alcun r_n (e la tesi è conseguentemente dimostrata),
- oppure prosegue all'infinito, determinando due successioni di reali, quella $a_0 < a_1 < \cdots$ crescente ma superiormente limitata da ciascuno dei b_i, e l'altra $b_0 > b_1 > \cdots$ decrescente ma inferiormente limitata da ciascuno degli a_i. Entrambe, allora, ammettono limite reale, rispettivamente a_∞ e b_∞, con $a_\infty \leq b_\infty$. Nessun reale tra a_∞ e b_∞ compresi può tuttavia appartenere alla successione di partenza. Infatti ogni singolo r_n, essendo minore di a_{n+1} o maggiore di b_{n+1}, è conseguentemente esterno ad a_∞ e b_∞. Dunque neppure in questo caso gli r_n esauriscono l'intervallo $]a, b[$.

Successive dimostrazioni, alcune dello stesso Cantor, ribadiranno la distinzione tra il numerabile e il continuo. Sono quelle che più facilmente si incontrano nei corsi universitari. Questa è tuttavia la prima da lui ideata. A onore del vero, all'argomento che tratta il caso algebrico, adoperando il concetto di altezza, contribuì in modo decisivo Dedekind, che fu spesso confidente di Cantor nelle sue ricerche. I due avevano elaborato nel 1872 due definizioni alternative ma equivalenti del concetto di numero reale, basandosi Cantor sulle così dette successioni di Cauchy di razionali, e Dedekind sulle sezioni in $\mathbb{Q}$. Per questa ragione erano venuti in contatto. Cantor aggiornò quindi Dedekind sulle sue ricerche a proposito sia della numerabilità dei reali algebrici che dell'innumerabilità di quelli trascendenti. Sul primo problema Dedekind gli fornì suggerimenti preziosi. Anzi la dimostrazione che alla fine Cantor pubblicò riprese quasi completamente quei consigli, semmai trascurandone uno – circa la possibilità di estendere l'argomento dai reali algebrici ai complessi algebrici, dimostrando anche nel loro caso la numerabilità. Cantor rimedierà a questa omissione solo alcuni anni dopo, e in modo abbastanza frettoloso. Del resto, Cantor dimenticò pure di ringraziare Dedekind per il suo aiuto: nel suo articolo non c'è segno alcuno di riconoscenza al collega. I motivi di questa apparente ingratitudine sono forse da cercare nella realtà scientifica tedesca di quegli anni e nella rivalità tra l'università di Berlino, che curava il *Journal für die reine und angewandte Mathematik* dove l'articolo apparve e che aveva avuto Cantor stesso come studente, e di Göttingen, dove Dedekind si era invece laureato. Cantor, dunque, avrebbe omesso il nome del suo interlocutore per non creare problemi diplomatici. A sua ulteriore scusante, si può aggiungere che la parte di gran lunga più originale e rivoluzionaria del lavoro è la seconda. Stupisce anzi che il titolo dell'articolo la trascuri del tutto, riservando la sua enfasi ai reali algebrici. Ma anche per questa stranezza c'è una spiegazione. Infatti a dirigere il *Journal für die reine und angewandte Mathematik* era in quel periodo anche Leopold Kronecker, che di Cantor era stato professore ma negli anni successivi si rivelò avversario critico e talora feroce, incapace di comprendere l'innovatività delle teorie sull'infinito. Il riferimento ai reali algebrici era dunque probabilmente solo una misura di prudenza, per ben disporre il maestro e distrarlo da ogni accenno all'argomento "scandaloso" delle cardinalità.

Bibliografia essenziale: [B.1] G. Cantor, Über eine Eigenschaft des Inbegriffes aller reellen algebraischen Zahlen, *Journal für die reine und angewandte Mathematik (Crelle)* 77 (1874), pp. 258-262 (è l'articolo considerato). [B.2] G. Cantor, *Gesammelte Abhandlungen mathematicschen und philosophischen Inhalts*, a cura di E. Zermelo, Springer, Berlin, 1932 (raccoglie tutte le opere di Cantor). [B.3] G. Cantor, *La formazione della teoria degli insiemi*, a cura di G. Rigamonti, Mimesis, Milano, 2012 (fornisce la traduzione italiana di vari lavori fondamentali di Cantor, incluso quello che ci interessa). [B.4] J. W. Dauben, *Georg Cantor. His Mathematics and Philosophy of the Infinite*, Princeton University Press, Princeton, 1990 (è la biografia più famosa di Cantor). [B.5] G. Lolli, *Nascita di un'idea matematica*, Edizione della Normale, Pisa, 2013 (tratta lungamente anche le teorie di Cantor). [B.6] C. Toffalori (a cura di), *Cantor. La teoria degli insiemi*, Grandangolo Scienza, Milano, Corriere della Sera, 2017 (è una breve presentazione di Cantor, della sua epoca e della sua opera, incluso l'articolo esaminato in questa nota).

[*] Sezione di Matematica – Scuola di Scienze e Tecnologie – Università degli Studi di Camerino

Zipf's «Rank-Size» Restatement

di Luciano Corso

[*Segue dal numero* 229]

Si ha:

$$H(x) = Pr(u \le x) = \int_{x_0}^{x} f(u)\,du = \int_{x_0}^{x} \frac{\alpha}{x_0} \cdot \left(\frac{x_0}{u}\right)^{\alpha+1} du =$$
$$= 1 - \left(\frac{x_0}{x}\right)^{\alpha} \qquad (2)$$

(il termine "Pr" sta per probabilità).

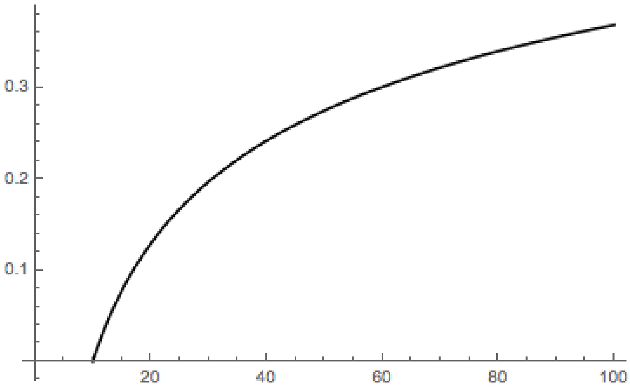


Fig.2. Il comportamento della distribuzione (cumulata delle frequenze) di Pareto con $\alpha = 0{,}2$, $x_0 = 10$. Qui, x è il reddito e $H(x)$ la frequenza cumulata dei redditieri. x_0 è il reddito di sopravvivenza.
Per $x \to +\infty, H(x) = 1$.

Il contributo di Zipf

Zipf modificò la funzione di Pareto e la ridusse alla forma:

$$f(A_r) = \frac{C}{r^a} \qquad \{r \mid r \in \mathbf{N_0}\} \qquad (3)$$

in cui $f(A_r)$ è la frequenza relativa (la funzione vale anche se si usa la frequenza assoluta) della parola A_r di rango r, C è una costante che spesso viene fatta corrispondere alla frequenza della parola di rango 1; C dipende dalla lunghezza del *corpus* e dal suo vocabolario, a è un indice inverso della ricchezza lessicale del *corpus*; più grande è a e più ristretto è il vocabolario del *corpus*; per Zipf, $a \approx 1$ in un *corpus* standard dove viene espressa una buona proprietà lessicale essenziale. La funzione cumulata di (3) è:

$$F(r \le k) = \sum_{r=1}^{k} f(A_r) = \sum_{r=1}^{k} \frac{C}{r^a} \quad con\ k \in \mathbf{N_0}, k \le \max(r). \qquad (4)$$

In linguistica, Zipf et al. si posero due questioni significative: 1) è vero che alla ricchezza del vocabolario di un testo corrisponde una maggiore qualità linguistica del testo e dell'autore? 2) In una comunicazione tra due soggetti se chi comunica usa un vocabolario ricco, chi riceve comprende meglio il comunicato ricevuto, a parità di *corpus*? Zipf e, prima di lui, Auerbach [B.3], dimostrano che nelle comunicazioni si tende a usare un linguaggio utilitaristico, essenziale e quindi sostanzialmente povero.

Se consideriamo una argomentazione composta di N parole, il linguaggio più povero è costituito da N parole uguali, quello più ricco da N parole diverse: nel primo caso o basta una sola parola per esprimere il concetto, e quindi la frase è ridondante, o siamo in presenza di un'argomentazione priva di significato; nel secondo caso o il testo è ben argomentato e chiaro o è privo di significato. Nel primo caso la funzione rango è uguale a $f(A_1) = N/N$ in $r = 1$. Se, invece, le N parole sono tutte diverse, il valore della funzione rango è: $f(A_r) = 1/N$ e $r = 1$ che si ripete N volte. Sono i due casi estremi della legge di Zipf; tutti gli altri si collocano tra questi due. Sperimentalmente è emerso che se $\alpha < 1$ significativamente, il vocabolario di chi comunica è ricco, se, invece, $\alpha > 1$ significativamente, il vocabolario del comunicatore è povero.

Applicazione: Applicando la riclassificazione dimensione rango di Zipf ai primi 54 articoli della Costituzione della Repubblica Italiana (Parte I: Diritti e doveri dei cittadini), dopo aver stabilito il significato di ciò che si deve intendere per parola, abbiamo ottenuto che il numero di parole è $N = 2872$, la parola "e" ha la massima frequenza assoluta $n_1 = 134$ cui corrisponde il rango $r = 1$; le parole distinte, con rango $r = 37$, sono 611. Non sono stati assegnati ranghi diversi a parole diverse con frequenza uguale, contrariamente a quanto fatto da certi autori [B.8]. Quindi, abbiamo stimato con il metodo dei minimi quadrati i parametri C e α di $f(A_r)$ considerando che per ogni r distinto abbiamo associata la frequenza relativa corrispondente e presa una sola volta. La (3) linearizzata diventa:

$$Ln\big(f(A_r)\big) = Ln(C) - \alpha \cdot Ln(r)\,. \qquad (5)$$

Applicando le derivate parziali alla funzione rispetto a $Ln(C)$ e ad α e considerando i dati sperimentali si ottiene:

$$\begin{cases} 37\,Ln(C) - 99.3306\,\alpha = -188.595 \\ -99.3306\,Ln(C) + 293.754\,\alpha = +\,535.852 \end{cases}$$

e risolvendo si ottiene: $Ln(C) \approx -2.16898, \alpha \approx +1.09073$.
Il modello adattato, quindi, risulta:

$f(A_r) \approx 0.114294/r^{1.09073}$.

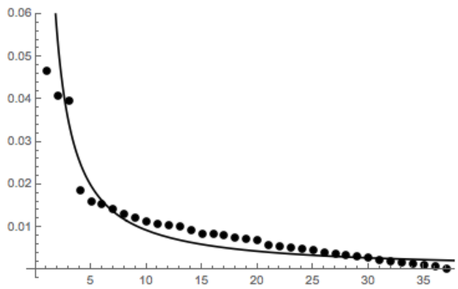


Fig.3. Il comportamento della distribuzione di Zipf nel caso dei 54 articoli della Costituzione della Repubblica italiana. Nel grafico r è sull' asse delle ascisse e $f(A_r)$ sull'asse delle ordinate.

Riferimenti Bibliografici: [B.1] G. K. Zipf, *Human Behaviour and the Principle of Least-Effort*, Addison-Wesley, Reading MA, (1949). [B.2] L. A. Adamic & B. A. Huberman in L. A. Adamic, B. A. Huberman, Zipf's law and the Internet, *Glottometrics* (2002), 3, 143-150. [B.3] M.E.J. Newman, Power laws, Pareto distribution and Zipf's law, *Contemporary Physics* (2006), 43, 323-351. [B.4] L. Vajani, *Statistica descrittiva*, ETAS libri, Milano, 1974. [B.5] L. Daboni, *Calcolo delle Probabilità ed elementi di Statistica*, UTET, Torino, 1980. [B.6] G. Landenna, D. Marasini, P. Ferrari, *Probabilità e Variabili Casuali*, ed. Il Mulino, Milano, 1997. [B.7] A. Lenci, *Parole e frequenze*, disponibile all'indirizzo http://webilc.ilc.cnr.it/~lenci/Appunti-LeggeZpf.pdf [Ultimo accesso 26/10/2017]. [B.8] R. H. Baayen, *Word Frequency Distributions*, Springer, 2001.