STATISTICA, anno LXIV, n. 2, 2004

IL PROBLEMA DELLA STIMA DA HERZEL A EFRON: L'EVOLUZIONE DEL PRINCIPIO DI CAMPIONAMENTO

Angela Montanari

1. L'inferenze statistica classica, da Fisher ad oggi, ha fondato la sua sottile rete di argomentazioni logiche, per giustificare il passaggio induttivo dal noto (il campione osservato) all'ignoto (la popolazione virtuale di riferimento), sul concetto di campione casuale inteso come esperimento casuale da cui generare per via assiomatica e analitica il supporto formale alla base di tutta la teoria statistica della stima. L'idea di variabile aleatoria campionaria o di distribuzione campionaria di una grandezza (o di una sintesi empirica di grandezze) hanno trovato nella dinamica combinatoria dell'universo dei campioni un punto di riferimento intuitivo e rassicurante.

Quasi tutti gli statistici del Novecento hanno finito per fare i conti con questo paradigma che consegnava alla metodologia statistica una possibile giustificazione dell'inferenza probabile.

Anche Amato Herzel si muoveva nell'alveo dell'inferenza statistica classica di impronta fisheriana e neymaniana e ha fondato gran parte della sua produzione scientifica proprio sulle suggestive potenzialità del concetto di distribuzione campionaria di cui ha sviluppato prevalentemente gli aspetti formali.

Un interesse molto forte per la statistica matematica pervade tutta l'opera di Amato Herzel. L'area in cui, forse più di altre, egli ha dimostrato le sue notevoli capacità analitiche è appunto lo studio delle distribuzioni teoriche e delle distribuzioni campionarie.

Nel corso delle sue ricerche, ha derivato le proprietà campionarie dello scostamento quadratico medio dalla mediana (Herzel, 1963), evidenziandone la distorsione nel finito e la correttezza asintotica. Ha dimostrato che nel campionamento con o senza reimmissione, per un'ampia classe di funzioni simmetriche, "le stime corrette con varianza minima possono essere ottenute come semplici trasformazioni da opportuni valori medi" e, partendo da questo risultato, è pervenuto a nuove espressioni delle stime corrette della varianza e dei momenti centrali di ordine superiore di funzioni campionarie (Herzel, 1982), fra cui la varianza (Herzel, 1985).

Gran parte dei suoi lavori hanno riguardato i parametri delle distribuzioni campionarie di alcuni indicatori statistici, introdotti nella letteratura statistica da Gini in un contesto puramente descrittivo, e trasferiti dagli sviluppi di Herzel in un più am-

pio scenario induttivo: media e varianza della distribuzione campionaria dell'indice di dissomiglianza (Herzel, 1963); media e varianza della distribuzione campionaria della probabilità di transvariazione, di cui ha derivato anche la legge distributiva sotto particolari ipotesi (Herzel, 1967); quarto momento e curtosi della distribuzione campionaria dell'indice di cograduazione (Herzel, 1972), e altro ancora.

Relazioni ricorsive, eleganti e rigorosi sviluppi formali si snodano nelle pagine dei suoi lavori, sempre accompagnati da una attenta riflessione sul loro significato nella letteratura statistica dei fenomeni empirici e sulle implicazioni logiche e interpretative che discendono da talune assunzioni.

Erano gli anni tra il 1960 e il 1970; di lì a poco la rivoluzione informatica, allora agli albori, avrebbe prodotto strumenti di elaborazione sempre più sofisticati, capaci di indagare fenomeni di elevata complessità attraverso la soluzione di sempre nuovi e stimolanti problemi di stima e avrebbe ad un tempo offerto gli strumenti e gli algoritmi per una rapida risposta a questi problemi, sostanzialmente velocizzando e ampliando quello che con un paziente lavoro analitico Herzel aveva già prefigurato.

2. L'idea vincente di variabile aleatoria campionaria e, ancor più, il concetto di universo dei campioni e di distribuzione campionaria hanno rappresentato per anni l'esperimento aleatorio che consentiva di generare modelli statistici e probabilistici. Per altro, così avevano fatto i probabilisti del Seicento e del Settecento quando simulavano i giochi di sorte per formulare i primi teoremi del calcolo delle probabilità. L'esperimento aleatorio costruito sui criteri probabilistici di campionamento da popolazioni reali o virtuali ha trovato la sua prima affermazione come potente strumento di simulazione durante la seconda guerra mondiale: un periodo che ha visto l'affermazione della statistica e dei suoi modelli in tanti settori strategici che richiedevano capacità di previsione e di decisione.

La simulazione, intesa come sperimentazioni su modelli, diventava così la base e la forza di molti sviluppi metodologici e avrebbe offerto alla statistica nuovi spunti di riflessione e ausili originali e semplici – e per questo talora abusati. Scrive Scardovi (1997) "Protagonista della simulazione su modelli formali è il calcolatore: laboratorio virtuale di esperimenti per mentem, in cui l'immaginazione, alleggerita del peso dei calcoli, resta libera di cimentarsi in tutti gli sviluppi ipotetico-deduttivi impliciti nel modello." E ancora "La teoria del campionamento casuale è il fondamento statistico del più noto e più usato criterio di simulazione mediante variabili casuali, il metodo di Monte Carlo: un metodo di generazione di variabili aleatorie attraverso distribuzioni campionarie, ispirato ai teoremi della convergenza stocatstica, sorto come strumento di indagine sul micromondo fisico e diffusosi poi a una grande varietà di contesti." (Scardovi, 1997).

Nota la legge distributiva di un carattere in una popolazione, attraverso il metodo Monte Carlo diventa immediato generare l'universo dei campioni e studiare in esso le proprietà della legge distributiva anche delle più complicate statistiche, funzione della *n*-pla campionaria. Al risultato formale si sostituisce un risultato numerico, la generalità delle formule cede il passo alla velocità dei calcoli; l'impatto di ogni variazione nella specificazione del modello trova immediata espressione nei valori che il calcolatore in pochi secondi rende disponibile.

Ma l'incalzare di nuovi problemi ha reso necessarie nuove soluzioni. Cosa fare, ad esempio, in assenza di informazioni preliminari adeguate sulla popolazione di riferimento?

All'esigenza, sempre più forte di determinare errori standard per particolari stimatori, di costruire intervalli di confidenza per parametri incogniti, di determinare *p*-values per certe statistiche test vera l'ipotesi nulla, sulla scorta delle sole informazioni contenute nel campione senza nessun altro supporto, risponde nel 1977 Bradly Efron proponendo nella famosa Ritz Lecture "Bootstrap methods: another look at the jackknife", che fu pubblicata 2 anni dopo, il metodo bootstrap (Efron, 1979). A circa 25 anni dalla sua nascita si contano circa 7.000 articoli che riportano sviluppi metodologici e applicazioni nei contesti empirici più disparati. Ma, "l'impatto del bootstrap ha trasceso sia la teoria che le applicazioni. Il bootstrap ha mostrato come l'uso della potenza del computer e dei calcoli iterativi possono giungere dove il calcolo teorico non può, e ciò introduce un diverso modo di pensare alla statistica. Non più la ricerca di soluzioni espresse in forma chiusa, ma l'impiego di algoritmi e iterazioni" (Casella, 2003).

Questo è il tipico contesto in cui collocare la soluzione bootstrap: un modello probabilistico P incognito, che dipende da un incognito vettore di parametri ha generato il vettore di dati osservati $\mathbf{x}$. Da $\mathbf{x}$ si calcola una statistica $\hat{\theta} = s(\mathbf{x})$ con l'obiettivo di stimare un parametro $\theta = t(P)$ (che può anche essere il prodotto di un processo molto complesso). Ciò che si vuole valutare è l'accuratezza di $\hat{\theta}$ per la stima di θ, in termini di distorsione, varianza, intervalli di confidenza, ecc.

La soluzione presuppone lo sviluppo del cosiddetto mondo bootstrap. Si costruisce una stima puntutale $\hat{P}$ di P da cui si generano nuovi vettori di dati bootstrap $\mathbf{x}^*$ e replicazioni bootstrap $\hat{\theta}^* = s(\mathbf{x}^*)$. Poiché $\hat{P}$ è completamente nota si può generare un numero arbitrario di $\hat{\theta}^*$ e utilizzare la variabilità delle repliche bootstrap per valutare l'accuratezza di $\hat{\theta}$. Si assume in questo modo che "la variabilità che si incontra campionando dal campione descriva quella che si sarebbe trovata campionando dalla popolazione" (Hall, 2002).

L'aspetto più importante e delicato del metodo sta nella possibilità di stimare P a partire da $\mathbf{x}$ *(plug-in-principle)*. E' l'unico momento induttivo del bootstrap -tutto il resto è semplice deduzione e calcolo - che lo configura come un modo per ricavare da una stima di P una misura di accuratezza per θ.

L'idea di ricampionare dal campione è molto vicina a quella del campionamento da una popolazione finita. Non sorprende quindi, come nota Hall (2002), che le radici del bootstrap possano essere rintracciate nelle ricerche sul campinoamento degli anni 40 e 50, e in particolare nei lavori di Mahalanobis (1946) in cui, nel metodo "*half sampling*", come suggerisce il nome i due campioni rappresentano uno il controllo dell'altro. "Il merito di Efron sta forse nell'aver combinato la potenza dell'approssimazione Monte Carlo con un ampia visione sul tipo di problemi che il bootstrap avrebbe potuto risolvere, in tal modo portando le prime idee del ricampionamento, fuori dall'ambito delle indagini campionarie, nel regno di una metodologia statistica universale" (Hall, 2002).

E come in un processo ciclico in cui ogni scoperta si fonda sulle conoscenze che la precedono e ad un tempo le amplia, così è fuori di dubbio che anche le indagini campionarie abbiano poi beneficiato del bootrstrap. (Lo documentano con chiarezza e dovizia di particolari, fra gli altri Shao (2002) e Lahiri (2002) celebrando i 25 anni di vita del bootrstrap.)

3. Il passaggio dal campione alla stima della densità di probabilità incognita che lo ha generato è semplice nel caso di un solo campione, $\mathbf{x} = x_1, x_2,...,x_n$, estratto mediante campionamento casuale semplice da una distribuzione di probabilità totalmente ignota. $\hat{P}$ può essere costruita assegnando probabilità $1/n$ a ciascun x_i. Ma non sempre i problemi ammettono soluzioni così semplici, e nonostante gli innumerevoli successi del bootstrap, ancora molti aspetti – non ultimo il principio di *plug in* – restano suscettibili di ulteriori sviluppi e approfondimenti.

Forse il calcolatore, che è stato il motore della "rivoluzione" bootstrap, rappresenterà ancora un partner insostituibile del ragionamento e della fantasia del ricercatore, che non dovrebbe tuttavia mai rinunciare, a fronte dei sorprendenti risultati numerici, all'"interpretazione del significato" investigativo di quei risultati. Un obiettivo che Herzel nei suoi lavori di statistica matematica non ha mai mancato di privilegiare, anche quando il momento tecnico sembrava prevalere. Infatti scriveva in un manuale didattico, a proposito del campionamento probabilistico: "Il grande vantaggio che esso presenta (...) è nella possibilità di costruire un modello matematico che usa certi risultati del calcolo delle probabilità e in particolare del calcolo combinatorio per porre su basi razionali e solide la scelta delle strategie possibili. Questo modello matematico, che costituisce la sostanza della teoria dei campioni, consente anche di valutare l'attendibilità dei risultati" (Herzel, 1994).

Dipartimento di Scienze Statistiche ANGELA MONTANARI
Università di Bologna

RIFERIMENTI BIBLIOGRAFICI

G. CASELLA, (2003), *Introduction to the Silver Anniversary of the Bootstrap*, "Statistical Science", 18, pp. 133-134.

B. EFRON, (1979), *Bootstrap methods: Another look at the jackknife*, "The Annals of Statistics", 7, pp. 1-26.

P. HALL, (2002), *A short prehistory of the bootstrap*, "Statistical Science", 18, pp. 158-167.

A. HERZEL, (1963), *Sui valori ordinali e sullo scostamento semplice medio dalla mediana nei campioni estratti con o senza ripetizione da popolazioni discrete e finite*, "Biblioteca del Metron", serie C: note e commenti, vol. II.

A. HERZEL, (1982), *On mean value and unbiased estimators in simple random sampling*, (1982), Statistica, XLII, no. 3.

A. HERZEL, (1985), *Campionamento semplice: stime della varianza della varianza campionaria*, "Statistica", XLV, no. 2.

A. HERZEL, (1994), *Il campionamento statistico,* in AAVV, "Metodi statistici per le scienze economiche e sociali", Monduzzi, Bologna.

P. LAHIRI, (2002), On the *impact of bootstrap in survey sampling and small-area estimation*, "Statistical Science", 18, pp. 199-210.
I. SCARDOVI, (1997), *Modelli di Simulazione*, "Estratto dal volume VII della Enciclopedia delle Scienze Sociali, Istituto della Enciclopedia Italiana, Giovanni Treccani".
J. SHAO, (2002), *Impact of the bootstrap on sample surveys*, "Statistical Science", 18, pp. 191-198.

RIASSUNTO

Il problema della stima da Herzel a Efron: l'evoluzione del principio di campionamento

L'articolo riprende i contributi di Amato Herzel sul problema della stima e sulla derivazione delle distribuzioni campionarie e riconosce, nel percorso che dagli sviluppi analitici di Herzel porta al bootstrap di Efron, un continuum governato dal principio di campionamento.

SUMMARY

The problem of statistical estimation from Herzel to Efron: the evolution of the sampling principle

The paper recalls Herzel's contributions to the development of statistical estimation theory and to the study of sampling distributions and recognizes in the path which starting from Herzel's algorithmic developments brings to Efron's bootstrap a continuum underlied by the sampling principle.