L'approccio alla statistica robusta basato sulla funzione d'influenza: appunti per un seminario

Matteo Pelagatti

6 aprile 2000

1 Introduzione e motivazione

La statistica parametrica, sia essa frequentista o bayesiana, ha come obiettivo primario quello di trovare procedure che, sotto un dato modello stocastico, siano ottime secondo qualche criterio. Tuttavia nulla viene detto sul comportamento di queste ultime, quando il modello ipotizzato è solo approssimativamente valido. L'obiettivo della statistica robusta è quello di predisporre strumenti per valutare la bontà delle procedure statistiche in intorni di modelli stocastici, e quindi di trovare procedure che mantengano buone proprietà anche quando il modello ipotizzato è solo un'approssimazione del "vero" modello. Vista in quest'ottica la statistica robusta può essere definita come *la statistica dei modelli parametrici approssimati*.

La necessità di una statistica robusta è brillantemente mostrata da Tukey [12] nel suo studio sulla distribuzione normale contaminata. Confrontando, infatti, stimatori di posizione e di scala applicati a campioni provenienti da una popolazione normale $N(0,1)$ e dalle popolazioni contaminate

$$(1-\eta)N(0,1)+\eta N(0,9) \quad \textit{Modello simmetrico} \tag{1}$$
$$(1-\eta)N(0,1)+\eta N(2,9) \quad \textit{Modello asimmetrico} \tag{2}$$

con $\eta \in (0,1)$, si giunge ai sorprendenti risultati riassunti nella tabella 1 e nella tabella 2. Se per la stima della media di una normale l'utilizzo della mediana campionaria in luogo della media campionaria porta ad una perdita d'efficienza del 36% è sufficiente il 10% di contaminazione nel modello (1) ed il 6% nel modello (2) per rendere le due stime ugualmente efficienti. Per quanto riguarda la stima del parametro di scala si nota che è sufficiente lo 0.2% di contaminazione nel modello (1) e lo 0.1% nel modello (2) per rendere la media degli scarti assoluti dalla media più efficiente della devizione standard campionaria (normalmente più efficiente del 12%).

Tukey commenta così i suoi risultati:

η	0.0000	0.0018	0.0282	0.1006
Var(media)	1.0000	1.0140	1.2252	1.8047
Var(mediana)	1.5708	1.5745	1.6315	*1.8047*
C.V.(deviazione standard)	0.7071	0.7627	1.1725	1.3540
C.V.(media scarti assoluti)	0.7555	*0.7627*	0.8514	0.9822
C.V.(mediana scarti assoluti)	1.1664	1.1668	*1.1725*	1.1899

Tabella 1: Modello simmetrico: $(1-\eta)N(0,1)+\eta N(0,9)$

η	0.0000	0.0008	0.0115	0.0617
Var(media)	1.0000	1.0097	1.1377	1.7254
Var(mediana)	1.5708	1.5727	1.5977	*1.7254*
C.V.(deviazione standard)	0.7071	0.7611	1.1694	1.5176
C.V.(media scarti assoluti)	0.7555	*0.7611*	0.8263	0.9973
C.V.(mediana scarti assoluti)	1.1664	1.1666	*1.1694*	1.1838

Tabella 2: Modello asimmetrico: $(1-\eta)N(0,1)+\eta N(2,9)$

> *A tacit hope in ignoring deviations from ideal models was that they would not matter; that statistical procedures which were optimal under the strict model would still be approximately optimal under approximate model. Unfortunately, it turned out that this hope was often drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians.*

Date queste premesse, ci si può legittimamente chiedere il motivo per cui non si possano risolvere i problemi fin'ora sollevati per mezzo di tecniche non parametriche. Quando il modello stocastico generatore dei dati è completamente sconosciuto, non esiste alternativa all'utilizzo della statistica parametrica; tuttavia, il più delle volte si possiede un modello che, per quanto approssimato, fornisce informazioni utili sul fenomeno in studio. Utilizzare la statistica non parametrica in una tale situazione, significa gettare nella spazzatura una notevole quantità d'informazione, mentre ricorrere alla statistica parametrica classica corrisponde un po' ad una passeggiata su un filo teso per aria.

Molto spesso gli statistici applicati si rendono conto della sensibiltà delle procedure classiche a deviazioni dal modello ipotizzato, ed in particolare alla presenza di *outliers* (valori anomali), e per far fronte al problema predispongono procedure ad hoc che solitamente si concretizzano nell'individuazione, per lo più artigianale, degli outliers e nella loro eliminazione o sostituzione. Tali pratiche si differenziano dalla filosofia che ispira i teorici della statistica robusta per il fatto che questi ultimi individuano procedure che godono di qualche criterio formale

di ottimalità e robustezza.

Se i semi della statistica robusta si possono rintracciare in molti scritti di statistici più o meno illustri del passato, i fondatori di una teoria della robustezza completa, formalizzata ed applicabile sono senz'altro Huber [6][7] e Hampel [2]. Gli approcci seguiti dai due autori non sono i medesimi: Huber segue un aproccio di tipo minimax, ovvero cerca di ottimizzare la situazione peggiore in cui ci si può trovare; Hampel è invece il padre dell'approccio infinitesimale, basato sulla funzione d'influenza. Nel presente scritto si parlerà esclusivamente di quest'ultimo, che secondo chi scrive risulta più intuitivo e più generalizzabile ad altri ambiti della statistica. Nonostante la differenza degli approcci, i risultati ottenuti dai due studiosi spesso coincidono.

2 L'approccio infinitesimale alla statistica robusta

L'approccio infinitesimale è basato su tre concetti fondamentali: *robustezza qualitativa*, *funzione d'influenza* e *punto di rottura*. Poiché molte statistiche dipendono solamente dalla funzione di ripartizione empirica (fre) dei dati (o dalla fre e dalla numerosità campionaria n), queste possono essere viste come funzionali definiti nello spazio delle distribuzioni di probabilità (o rimpiazzate da funzionali per ogni n), rendendo possibile l'applicazione di concetti quali continuità e derivazione.

La robustezza qualitativa è definita come equicontinuità delle distribuzioni della statistica al variare di n. La robustezza qualitativa può essere considerata una condizione di robustezza necessaria, benché piuttosto debole: permette già di eliminare un grosso numero di procedure classiche, non fornendo tuttavia alcuna informazione sulle differenze tra le procedure qualitativamente robuste.

Lo strumento più ricco e foriero d'informazioni dell'approccio infinitesimale è la funzione d'influenza e le varie quantità da esse derivate. Essa descrive l'effetto (approssimato e standardizzato) di un'osservazione aggiuntiva di valore x su una statistica T, dato un ampio campione estratto dalla distribuzione F. Rimandando ai prossimi paragrafi una definizione più esaustiva, la funzione d'influenza $\mathrm{IF}(x; T, F)$ può essere descritta come la derivata direzionale in F della statistica T nella direzione della distribuzione δ_x di Dirac. A questi punto è possibile approssimare localmente la statistica T con il suo sviluppo di Taylor fermato al primo termine. Poiché si sta assumendo che la vera distribuzione della popolazione sia in un intorno di quella ipotizzata, e dato che per n sufficientemente grande la funzione di ripartizione empirica sarà simile a quella della popolazione, si può studiare il comportamento di T in un intorno della distribuzione ipotizzata per mezzo della funzione d'influenza.

Dato che la funzione d'influenza consente uno studio solamente locale del

comportamento di una statistica, è necessario affiancarle un uno strumento che misuri la robustezza in modo globale. Il punto di rottura assolve proprio a questo scopo. Informalmente esso può essere definito come la più piccola frazione di osservazioni (anomale) che possono portare il valore della statistica oltre ogni limite.

Usando una brillante similitudine di Huber [8], si può vedere la statistica T come un ponte: la robustezza qualitativa richiede che la stabilità del ponte debba essere poco inficiata da piccole perturbazioni, la funzione d'influenza misura gli effetti di perturbazioni infinitesime ed il punto di rottura indica quanto possono essere grandi le perturbazioni prima che il ponte ceda.

3 Ipotesi e notazione

Nei prossimi paragrafi si farà spesso riferimento ad un campione di osservazioni $(x_1, x_2, \dots, x_n)$, realizzazioni delle variabili casuali reali $(X_1, X_2, \dots, X_n)$ i.i.d.[1], definite su un medesimo spazio di probabilità $(\Omega, \mathcal{B})$. Si indichrà con $\mathcal{F}(\Omega)$ l'insieme di tutte le misure di probabilità definite sullo spazio $(\Omega, \mathcal{B})$, e si userà una lettera latina maiuscola sia per la misura di probabilità, sia per la funzione di ripartizione (per es. $F(A)$ e $F(x)$). θ sarà il parametro incognito della distribuzione, appartenente allo spazio parametrico $\Theta \subseteq \mathbb{R}$.

Si prenderanno in considerazione solo statistiche (i) che hanno una rappresentazione come funzionale: $T_n(G_n) = T(G_n)\ \forall n, G_n$, dove $T : (\subseteq \mathcal{F}(\Omega)) \to \mathbb{R}$ e G_n è la fre , o (ii) che possono essere asintoticamente sostituite da funzionali, cioè tali che esiste un funzionale T per cui vale

$$T_n(X_1, \dots, X_n) \xrightarrow[n\to\infty]{p} T(G) \tag{3}$$

quando le osservazioni sono i.i.d. con distribuzione G. Utilizzando i funzionali, la nozione di consistenza assume la forma

$$T(F_\theta) = \theta \quad \forall \theta \in \Theta, \tag{4}$$

dove T è stimatore di θ.

Si noti che molti degli stimatori classici possono essere scritti sotto forma di funzionali: la media campionaria, per esempio, si può scrivere

$$\bar{X} = \int x F_n(dx), \tag{5}$$

dove

$$F_n = n^{-1} \sum_1^n \delta_{x_i} \tag{6}$$

[1]Non si tratterà in questa sede della robustezza rispetto all'assunzione di indipendenza e uguale distribuzione.

è la distribuzione empirica del campione $(x_1, \dots, x_n)$, con δ_x misura di Dirac, che assegna massa 1 al punto x; gli stimatori di massima verosimiglianza hanno forma

$$\widehat{\theta} = \left\{ \theta \in \Theta : \int \log f(x, \theta) F_n(dx) = \max \right\}. \tag{7}$$

Il simbolo $\mathcal{L}_G(T)$ verrà usato col significato "legge di distribuzoine della statistica T sotto G".

4 Robustezza qualitativa

Nelle righe precedenti si è spesso utilizzato il concetto di vicinanza tra disribuzioni di probabilità e di intorno di una distribuzione F. Per formalizzare tale nozione è necessario definire una metrica nello spazio delle distribuzioni di probabilità, ma affinché la distanza utilizzata abbia un significato pratico bisogna dare una migliore definizione del concetto di "moderata deviazione dal modello parametrico ipotizzato". Hampel [3] cita tre principali motivi per cui la distribuzione reale delle osservazioni devi dal modello teorico ipotizzato:

1. l'arrotondamento od il raggruppamento in classi,
2. l'occorrenza di valori anomali,
3. il modello teorico pensato già in partenza come approssimazione (magari in virtù di qualche teorema centrale limite).

Pur non essendoci unanimità sulla scelta della metrica più adatta, Hampel propone la distanza di Prohorov, che oltre a cogliere tutti e tre i tipi di deviazione enunciati, consente il confronto anche tra distribuzioni assolutamente continue e distribuzioni discrete.

Definizione 1 (Distanza di Prohorov) [2] *Sia $(\Omega, \mathcal{B})$ uno spazio di probabilità dove Ω è uno spazio metrico (con distanza d) completo e separabile, e $\mathcal{B}$ è la relativa σ-algebra di Borel. Ad ogni $A \subset \Omega$, $A \in \mathcal{B}$ si associ l'intorno*

$$A^\varepsilon := \{x \in \Omega : \inf_{y \in A} d(x, y) < \varepsilon\sigma\}, \tag{8}$$

dove σ è una fattore di scala introdotto per rendere la distanza di Prohorov invariante rispetto a trasformazioni lineari dello spazio Ω. Siano F e G due misure di probabiltià su $(\Omega, \mathcal{B})$. La distanza di Prohorov tra tra F e G è data da

$$\pi(F, G) := \inf\{\varepsilon : F(A) \leq G(A^\varepsilon) + \varepsilon, \quad \forall A \in \mathcal{B}\}. \tag{9}$$

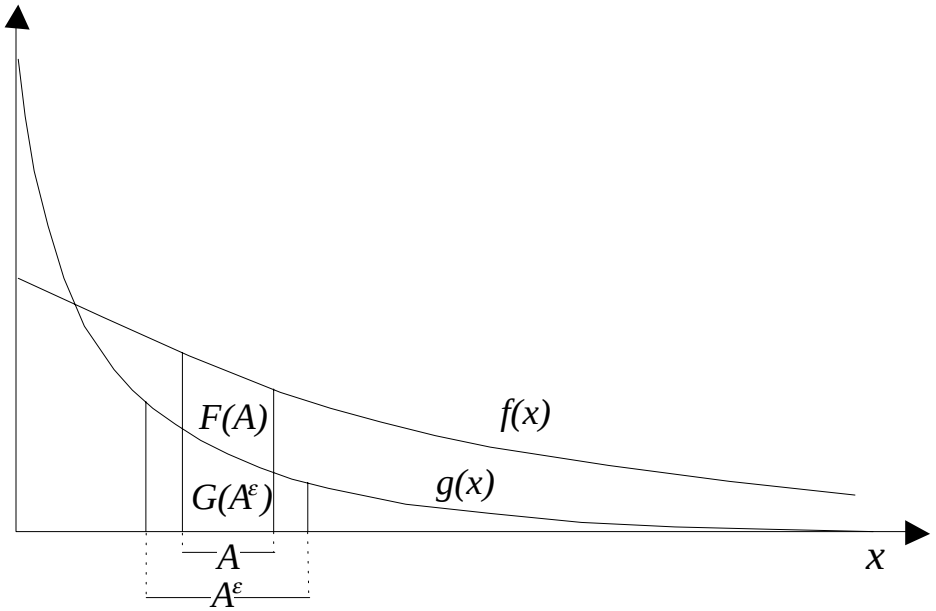

Figura 1: Elementi presenti nella definizione della distanza di Prohorov

In figura 1 sono rappresentati gli elementi presenti nella definizione della distanza di Prohorov. Non è difficile vedere che $0 \leq \pi(F, G) \leq 1$. Si può dare una definizione più esplicita della distanza Prohorov:

$$\pi(F, G) = \max_{A}\{\rho(A)\} \tag{10}$$

dove

$$\rho(A) = \min_{\varepsilon}\{\varepsilon : F(A) \leq G(A^{\varepsilon}) + \varepsilon\}. \tag{11}$$

Osservazioni. Quando F e G hanno densità limitate, la diseguaglianza nella definizione di $\pi(F, G)$ si può sostituire con un'uguaglianza. Inoltre si può dimostrare che la (10) implica l'ugluaglianza dei valori delle funzioni di densità f e g sulle frontiere degli insiemi A e A^{ε}:

$$f[\mathrm{Fr}(A_0)] = g[\mathrm{Fr}(A_0^{\varepsilon})]. \tag{12}$$

Quest'ultima relazione è di grande aiuto nella ricerca dell'insieme ottimale A_0.

Il primo fondamentale concetto di robustezza, nell'approccio infinitesimale, è quello di Robustezza qualitativa. Una statistica T è qualitativamente robusta se moderate deviazioni della distribuzione ipotizzata comportano solamente moderate deviazioni della legge di T, per ogni ampiezza campionaria.

[2]Si da una definizione della distanza di Prohorov valida per spazi di probabilità e leggermente modificata per mezzo dell'aggiunta di un fattore di scala.

Definizione 2 (Robustezza qualitativa) *Sia T_n una statistica, funzione di un campione di n osservazioni estratte indipendentemente dalla medesima distribuzione F. T_n è robusto in F se, dato $\epsilon > 0$, esiste un $\delta > 0$ tale che*

$$\begin{aligned} &\forall G \in \mathcal{F}(\Omega) \quad \forall n \in \mathbb{N} \\ &\pi(F, G) < \delta \Rightarrow \pi(\mathcal{L}_F(T_n), \mathcal{L}_G(T_n)) < \epsilon. \end{aligned} \tag{13}$$

La statistica T_n si dice (ovunque) robusta se la (13) *vale per ogni $F \in \mathcal{F}(\Omega)$.*

Esempio 1 (Non robustezza della media campionaria) *Si vole verificare la robustezza della statistica media campionaria quando si ipotizza che la popolazione abbia densità $f(x) = N(\mu, \sigma^2)$, mentre in realtà è distribuita secondo $g(x) = (1-\eta)N(\mu, \sigma^2) + \eta N(\alpha, \sigma^2)$ con $\eta \in (0, 1)$. È noto che $\mathcal{L}_F(\bar{X}) = N(\mu, \sigma^2/n)$, mentre si può mostrare che*

$$\mathcal{L}_G(\bar{X}) = \sum_{i=0}^{n} \binom{n}{i} \eta^i (1-\eta)^{n-i} N\left(\mu + \frac{i(\alpha-\mu)}{n}, \frac{\sigma^2}{n}\right). \tag{14}$$

Se nella (8) *si pone il parametro di scala uguale alla deviazione standard di F, utilizzando la* (12)*, si trova che l'insieme ottimale A_0 per calcolare la distanza tra F e G è approssimativamente*

$$A_0 = \left[\frac{\alpha + \mu + \sigma}{2} - \frac{\sigma^2 \log n}{\alpha - \mu}, \infty\right] \tag{15}$$

Dalla (10) *si ottiene $F(A_0) \approx \eta$, $G(A_0^\varepsilon) \approx 0$ e quindi $\pi(F, G) \approx \eta$. Analogamente si ottiene $\pi[\mathcal{L}_F(\bar{X}), \mathcal{L}_G(\bar{X})] \approx 1 - (1-\eta)^n$. Dato che un η piccolo non implica necessariamente un ϵ piccolo, la definizione di robustezza qualitativa non è soddisfatta.*

Purtroppo la proprietà di robustezza qualitativa non è facilmente verificabile ed inoltre iduce solo una dicotomia tra statistiche robuste e statistiche non robuste senza dare una misura della robustezza.

5 La funzione d'influenza

Lo strumento più ricco e applicabile dell'approccio infinitesimale è senza dubbio la funzione d'influenza (IF) con tutte le quantità che da essa derivano, introdotta da Hampel [2][4] inizialmente col nome di curva d'influenza. La IF è essenzialmente uno strumento euristico che gode di una importante interpretazione intuitiva.

Sia T una statistica (funzionale) definita su un sottoinsieme convesso di $\mathcal{F}(\Omega)$, si dirà che T è Gâteaux differenziabile in F, se esiste una funzione reale a_1 tale che per ogni G nel dominio di T vale

$$\lim_{t \to} \frac{T((1-t)F + tG) - T(F)}{t} = \int a_1(x) G(dx) \tag{16}$$

o equivalentemente

$$\frac{\partial}{\partial t}[T((1-t)F+tG)-T(F)]_{t=0}=\int a_1(x)G(dx). \tag{17}$$

Ponendo nella (17) $G=F$ è chiaro che

$$\int a_1(x)F(dx)=0 \tag{18}$$

da cui

$$\begin{aligned}\frac{\partial}{\partial t}[T((1-t)F+tG)-T(F)]_{t=0}&=\int a_1(x)dG(dx)=\\&=\int a_1(x)(G-F)(dx).\end{aligned} \tag{19}$$

Essendo $a_1(x)$ definita solo implicitamente nella (17), non è ancora chiaro il suo significato. Se si sostituisce G con il δ_x di Dirac, e quest'ultimo è nel dominio di T, si può dare una formulazione più debole ed esplicita della (17).

Definizione 3 (Funzione d'influenza) *La funzione d'influenza di T in F è data da*

$$\mathrm{IF}(x;T,F):=\lim_{t\downarrow 0}\frac{T((1-t)F+t\delta_x)-T(F)}{t} \tag{20}$$

per tutti i valori di $x\in\Omega$ per cui il limite esiste.

Si noti che la IF può essere equivalentemente definita con

$$\mathrm{IF}(x;T,F):=\frac{\partial}{\partial t}[T((1-t)F+t\delta_x]_{t=0}. \tag{21}$$

Le condizioni d'esistenza della funzione d'influenza sussistono praticamente in tutte le applicazioni statistiche comuni, per cui non è in genere necessario andarle a verificarle.

Esistono versioni finite della IF, che aiutano a dare una interpretazione alla IF stessa. La *sensistivity curve* (SC) di Tukey [13]

$$\mathrm{SC}_{n-1}(x):=n[T_n(x_1,\dots,x_{n-1},x)-T_{n-1}(x_1,\dots,x_{n-1})], \tag{22}$$

misura l'effetto di un'osservazione aggiuntiva x su una statistica T_n in funzione del valore di x. Se la statistica T_n è un funzionale, cioè se $T_n(x_1,\dots,x_n)=T(F_n)$ per ogni n, allora

$$\mathrm{SC}_{n-1}(x)=\frac{\left[T\left(\left(1-\frac{1}{n}\right)F_{n-1}+\frac{1}{n}\delta_x\right)-T(F_{n-1})\right]}{\frac{1}{n}}. \tag{23}$$

Quest'ultima espressione è un caso particolare della (20), dove F_{n-1} è un'approssimazione di F e $t = 1/n$. In molte situazioni[3] $\mathrm{SC}_{n-1}(x)$ covergerà a $\mathrm{IF}(x; T, F)$.

Se G è una distribuzione "vicina" a F, esiste uno sviluppo[4] (di von Mises) del funzionale T, analogo a quello di Taylor, che, fermato al primo ordine, è dato da

$$\begin{aligned} T(G) &= T(F) + \int \mathrm{IF}(x; T, F)(G - F)(dx) + \text{resto} = \\ &= T(F) + \int \mathrm{IF}(x; T, F)G(dx) + \text{resto}. \end{aligned} \tag{24}$$

Quando le osservazioni del campione sono realizzazioni di variabili casuali $X_i \sim$ i.i.d. con distribuzione G, per il teorema di Glivenko–Cantelli[5] $G_n \to G$ al crescere di n. Quindi, per n sufficientemente grande è possibile sostituire G con G_n nella (24):

$$\begin{aligned} T(G_n) &\approx T(F) + \int \mathrm{IF}(x; T, F)G_n(dx) + \text{resto} = \\ &= T(F) + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IF}(x_i; T, F) + \text{resto}. \end{aligned} \tag{25}$$

Prendendo G_n, e quindi anche X_i e $T(G_n)$, come variabili casuali, e con un banalissimo passagio algebrico si ottiene

$$\sqrt{n}(T(G_n) - T(F)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathrm{IF}(X_i; T, F) + \text{resto}. \tag{26}$$

Il primo termine dopo l'uguale è asintoticamente normale in virtù del teorema centrale limite. Sotto condizioni di norma vere nelle applicazioni statistiche, il resto è trascurabile per $n \to \infty$, cosicché

$$\sqrt{n}(T(G_n) - T(F)) \sim N(0, V(T, F)), \quad \text{quando } n \to \infty \tag{27}$$

con

$$V(T, F) = \int \mathrm{IF}(x; T, F)^2 F(dx). \tag{28}$$

Vi sono alcune quantità derivate dalla IF, che ne riassumono gli aspetti rilevanti per quanto riguarda lo studio della robustezza di una statistica, e che originano altrettante nozioni di robustezza. La più importante è la *gross-error sensitivity*, che misura la sensibilità di una statistica alla presenza di outliers.

[3] ma non sempre: il problema risiede nella sostituzione di F con F_{n-1}

[4] Per una trattazione matematicamente rigorosa di quanto si sta per esporre si veda [14][1].

[5] (**Teorema di Glivenko–Cantelli**) Siano le v.c. $X_1, X_2, \ldots, X_n$ i.i.d. con distribuzione F e sia F_n la relativa funzione di distribuzione empirica. Posto $D_n = \sup_x |F_n(x) - F(x)|$, $D_n \to 0$ con probabilità 1, quando $n \to \infty$

Definizione 4 (Gross-error sensitivity) *La gross-error sensitivity della statistica T sotto la distribuzione F è data da*

$$\gamma^* := \sup_x |\mathrm{IF}(x; T, F)|, \tag{29}$$

dove il dominio di x è costituito da tutti i punti di Ω in cui la IF *esiste.*

Alla gross-error sensitivity è legata la nozione di *B-robustezza* (dove B sta per *bias*), che può essere vista come robustezza rispetto agli outliers.

Definizione 5 (B-robustezza) *La statistica T è B-robusta sotto la distribuzione F se $\gamma^* < \infty$.*

La seconda quantità derivata dalla IF è la *local-shift sensistivity*, che misura la sensibilità di una statistica ad approssimazioni dovute agli arrotondamenti dei valori delle osservazioni.

Definizione 6 (Local-shift sensitivity) *La local-shift sensitivity della statistica T sotto la distribuzione F è data da*

$$\lambda^* := \sup_{x \neq y} \frac{|\mathrm{IF}(y; T, F) - \mathrm{IF}(x; T, F,)|}{|y - x|}, \tag{30}$$

dove il dominio di x e di y è costituito da tutti i punti di Ω in cui la IF *esiste.*

In realtà la local-shift sensitivity non gioca un grosso ruolo nello studio della robustezza di una statistica , anche perché è facile notare che quando la IF ha un salto, $\lambda^* = \infty$, indipendentemente dalla dimensione del salto. È comunque desiderabile che la local-shift sensitivity sia finita.

Il rigetto dei valori estremi si palesa nella IF con l'annullamento della stessa per valori "grandi" (in modulo) della x.

Definizione 7 (Rejection point) *Il rejection point dello stimatore T sotto la distribuzione F è definito da*

$$\rho^* := \inf\{r > 0 : \mathrm{IF}(x; T, F) = 0 \quad \textit{quando } |x - \mu| > r\}, \tag{31}$$

dove μ è un opportuno centro della distribuzione F.

La finitezza di ρ^* è una proprietà desiderabile di una statistica. Le statistiche che godono di tale proprietà si dicono *redescending*.

6 Il punto di rottura

Come complemento delle definizioni locali di robustezza date nei paragrafi precedenti, s'introduce ora la nozione di punto di rottura (*breakdown point*), che fornisce invece una misura globale di robustezza. Il punto di rottura della statistica T, stimatore del parametro θ, è il valore della distanza (di Prohorov) dalla distribuzione teorica F_θ oltre alla quale T fornisce valori arbitrariamente lontani dal valore di θ.

Definizione 8 (Punto di rottura) *Il punto di rottura δ^* della sequenza di stimatori $\{T_n\}_{n\geq 1}$ sotto la distribuzione F è definito da*

$$\delta^* = \delta^*(T_n, F) := \sup\big\{\delta \leq 1 : \exists \text{ un compatto } K_\delta \subset \Theta \text{ tale che } \pi(F,G) < \delta \Rightarrow G(\{T_n \in K_\delta\}) \xrightarrow{n\to\infty} 1\big\}. \quad (32)$$

Quando $\Theta = \mathbb{R}$ la (32) diventa

$$\delta^* = \delta^*(T_n, F) := \sup\big\{\delta \leq 1 : \exists\, r_\delta \in \mathbb{R}^+ \text{ tale che } \pi(F,G) < \delta \Rightarrow G(\{|T_n| \leq r_\delta\}) \xrightarrow{n\to\infty} 1\big\}. \quad (33)$$

Benché nella definizione δ^* dipenda dalla distribuzione F, in pratica spesso ciò non avviene.

Esistono diverse definizioni del punto di rottura. La definizione qui proposta è quella di Hampel [3], per definizioni alternative si consulti per esempio Huber [9] e Rey [11].

Vi sono versioni finite anche del punto di rottura. La seguente definizione è valida per stimatori di parametri di locazione.

Definizione 9 (Punto di rottura finito) *Il punto di rottura finito dello stimatore T_n dato il campione $(x_1, x_2, \dots, x_n)$ è definito da*

$$\delta_n^*(T_n; x_1, \dots, x_n) := \frac{1}{n} \min\left\{ m : \max_{i_1,\dots,i_m} \sup_{y_1,\dots,y_m} |T_n(z_1, \dots, z_n)| = \infty \right\}, \quad (34)$$

dove la n-upla $(z_1, z_2, \dots, z_n)$ si ottiene sostituendo agli m valori campionari $x_{i_1}, \dots, x_{i_m}$ i valori arbitrari $y_1, \dots, y_m$.

In genere δ_n^* non dipende dai valori della n-upla campionaria mentre dipende debolmente da n. In molte situazioni comuni $\lim_{n\to\infty} \delta_n^* = \delta^*$. L'interpretazione del punto di rottura finito è elementare: δ_n^* è la più piccola frazione di valori anomali che può portare il valore dello stimatore oltre ogni limite[6].

[6] A volte (per es. quando il supporto di X ed il codominio di T_n sono finiti) ha più senso nell'ultima parte della (34) sostituire $s_1 > T_n(z_1, \dots, z_n) > s_2$ dove (s_1, s_2) è un range ragionevole per il fenomeno in studio

7 Proprietà di alcuni noti stimatori di posizione

Sia $F_\theta(x) = \Phi(x-\theta)$, dove Φ è la funzione di ripartizione di una normale standard. Sia inoltre θ_0 il vero valore del parametro da stimare. Si vuole valutare la IF della media aritmetica $\bar{X} = \int x\Phi(dx)$ sotto le ipotesi premesse.

$$\begin{aligned}
\mathrm{IF}(x; \bar{X}, \Phi) &= \lim_{t\downarrow 0} \frac{\int u[(1-t)\Phi + t\delta_x](du) - \int u\Phi(du)}{t} = \\
&= \lim_{t\downarrow 0} \frac{(1-t)\int u\Phi(du) + t\int u\delta_x(du) - \theta_0}{t} = \\
&= \lim_{t\downarrow 0} \frac{\theta_0 - t\theta_0 + tx - \theta_0}{t} = \\
&= \lim_{t\downarrow 0} \frac{t(x-\theta_0)}{t} = \\
&= x - \theta_0
\end{aligned} \tag{35}$$

Dalla IF si ricava $\gamma^* = \infty$, $\lambda^* = 1$, $\rho^* = \infty$ e quindi la media non è B-robusta sotto la distribuzione normale; inoltre il punto di rottura finito è $1/n$ e quello asinotoico è 0.

Sia $\Omega = \mathbb{N} \cup \{0\}$, $\Theta = (0, \infty)$ e F_θ una distribuzione di Poisson, con funzione di probabilità

$$f_\theta(x) = \frac{\theta^x e^{-\theta}}{x!} \tag{36}$$

Si indichi ancora con θ_0 il vero valore del parametro da stimare. Lo stimatore di massima verosimiglianza di θ è la media campionaria

$$\bar{X} = T(F) = \int_\Omega uF(du) = \int_\Omega uf(u)d\lambda(u) = \sum_{x=0}^{\infty} xf(x), \tag{37}$$

dove λ è la misura conteggio. La IF è data da

$$\begin{aligned}
\mathrm{IF}(x; \bar{X}, F_{\theta_0}) &= \lim_{t\downarrow 0} \frac{\sum_{k=0}^{\infty} k[(1-t)f_{\theta_0}(k) + tI_x(k)] - \sum_{k=0}^{\infty} kf_{\theta_0}(k)}{t} = \\
&= \lim_{t\downarrow 0} \frac{t\sum_{k=0}^{\infty} kI_x(k) - t\sum_{k=0}^{\infty} kf_{\theta_0}(k)}{t} = \\
&= x - \theta_0 \qquad \forall x \in \mathbb{N} \cup \{0\}
\end{aligned} \tag{38}$$

Sia $m(G) = G^{-1}(1/2)$ la mediana della distribuzione G, che in caso di non unicità viene posta uguale alla semisomma degli estremi dell'intervallo $\{t; G(t) =$

$1/2\}$, e nel caso in cui G abbia nel punto t un salto per cui $G(t^-) < 1/2$ e $G(t) > 1/2$, viene posta uguale a t. Sia F una distribuzione di probabilità assolutamente continua, con densità limitata f e mediana m_0 unica. Si vuole ricavare la IF della mediana $m(F)$. Dalla definizione di IF si ricava:

$$\begin{aligned} \mathrm{IF}(x; m, F) &= \lim_{t \downarrow 0} \frac{m[(1-t)F + t\Delta_x] - m(F)}{t} = \\ &= \frac{m_0 + \alpha_t - m_0}{t} = \frac{\alpha_t}{t}, \end{aligned} \tag{39}$$

dove Δ_x è la funzione di distribuzione relativa alla misura di Dirac, e dove si è posto

$$m[(1-t)F + t\Delta_x] = m_0 + \alpha_t. \tag{40}$$

Per ovvie ragioni varrà

$$\begin{cases} \alpha_t \geq 0 & \text{se } x > m_0 \quad \Rightarrow \mathrm{IF}(x) \geq 0 \\ \alpha_t = 0 & \text{se } x = m_0 \quad \Rightarrow \mathrm{IF}(x) = 0 \\ \alpha_t \leq 0 & \text{se } x < m_0 \quad \Rightarrow \mathrm{IF}(x) \leq 0 \end{cases}. \tag{41}$$

Da quanto premesso si può scrivere

$$\int_{-\infty}^{m_0+\alpha_t} d[(1-t)F + t\delta_x] = \frac{1}{2}. \tag{42}$$

Sia $x > m$ e quindi $\alpha_t > 0$,

$$(1-t)\int_{-\infty}^{m_0} dF + (1-t)\int_{m_0}^{m_0+\alpha_t} dF + t\int_{-\infty}^{m_0} d\delta_x + t\int_{m_0}^{m_0+\alpha_t} d\delta_x = \frac{1}{2}, \tag{43}$$

Il primo integrale è pari a $1/2$, essendo m_0 la mediana di F. Il terzo integrale è nullo dato che per ipotesi $x > m_0$. Semplificando si ottiene

$$-\frac{t}{2} + (1-t)\int_{m_0}^{m_0+\alpha_t} dF + t\int_{m_0}^{m_0+\alpha_t} d\delta_x = 0. \tag{44}$$

Si vuole ora dimostrare che il secondo integrale, per t sufficientemente piccolo, è nullo in quanto vale $x > m_0 + \alpha_t$; si vuole cioè dimostrare che

$$\int_{-\infty}^{x} d[(1-t)F + t\delta_x] > \frac{1}{2} \quad \Rightarrow \quad m_0 + \alpha_t < x. \tag{45}$$

Per far questo sia $0 < \epsilon < x - m_0$, e quindi $m_0 < x - \epsilon$, e si calcoli

$$\begin{aligned} [(1-t)F + t\Delta](x-\epsilon) &= (1-t)\int_{-\infty}^{x-\epsilon} dF + t\int_{-\infty}^{x-\epsilon} d\delta_x = \\ &= (1-t)F(x-\epsilon) \xrightarrow{t \downarrow 0} F(x-\epsilon) > \frac{1}{2}; \end{aligned} \tag{46}$$

ciò implica che $[(1-t)F + t\Delta](x-\epsilon) > 1/2$ definitivamente (cioè $\forall 0 < t < t_0$), e quindi anche $m[(1-t)F + t\Delta] = m_0 + \alpha_t < x - \epsilon < x$, da cui segue la tesi.

Tornando alla (45), per il teorema della media integrale, il primo integrale vale

$$\int_{m_0}^{m_0+\alpha_t} dF = \int_{m_0}^{m_0+\alpha_t} f(u)du = \alpha_t f(\theta) \quad \text{con } \theta \in [m, m+\alpha_t]. \tag{47}$$

Per quanto mostrato fino a qui, per $0 < t < t_0$ e $x < m_0$, la (45) si può riscrivere come segue:

$$\begin{aligned} -\frac{t}{2} + (1-t)\alpha_t f(\theta) &= 0 \\ \alpha_t = \frac{t}{2(1-t)f(\theta)} &\asymp_{t\downarrow 0} \frac{t}{2f(m_0)}, \end{aligned} \tag{48}$$

e la funzione d'influenza vale, per $x > m_0$,

$$\lim_{t\downarrow 0} \frac{\alpha_t}{t} = \lim_{t\downarrow 0} \frac{t}{2tf(m_0)} = \frac{1}{2f(m_0)}. \tag{49}$$

Operando in modo analogo per $x < m_0$, si ottiene la IF della mediana per ogni valore di x:

$$\mathrm{IF}(x; m, F) = \frac{\mathrm{sgn}(x - m_0)}{2f(m_0)}. \tag{50}$$

8 Stimatori M

Gli stimatori M sono una generalizzazione degli stimatori di massima verosimiglianza proposta da Huber [6].

Definizione 10 (Stimatori M) *Sia $\rho : (\Omega, \Theta) \to \mathbb{R}$, una funzione dotata di derivata $\psi(x, \theta) = (\partial/\partial\theta)\rho(x, \theta)$. Uno stimatore M è definito da quel T_n per cui*

$$\sum_{i=1}^{n} \rho(X_i, T_n) = \min_{T_n}!, \tag{51}$$

o equivalentemente (...)

$$\sum_{i=1}^{n} \psi(X_i, T_n) = 0. \tag{52}$$

Se nella (51) si pone $\rho(X_i, T_n) = -\ln f_{T_n}(X_i)$, si ottiene lo stimatore di massima verosimiglianza. Se G_n è la funzione di distribuzione empirica del campione, la (52) può essere scritta $T(G_n)$, dove T è il funzionale definito implicitamente da

$$\int \psi(x, T(G))dG(x) = 0. \tag{53}$$

Per calcoloare la IF degli stimatori M, nella (53) si sostituisca G con $F_t = (1 - t)F + t\Delta_x$ e si derivi rispetto a t, posto che $\psi'(x, \theta) = (\partial/\partial\theta)\psi(x, \theta)$ esista in $T(F)$ e che l'integrazione e la derivazione siano scambialbili.

$$\begin{aligned}
0 &= \frac{\partial}{\partial t}\left\{ \int \psi(y, T(F_t))d[(1-t)F + t\delta_x] \right\}_{t=0} = \\
&= \frac{\partial}{\partial t}\left\{ \int \psi(y, T(F_t))dF - t\int \psi(y, T(F_t))dF + +t\int \psi(y, T(F_t))d\delta_x \right\}_{t=0} = \\
&= \left[\frac{\partial T(F_t)}{\partial t}\right]_{t=0} \int \psi'(y, T(F))dF - \left[\frac{\partial T(F_t)}{\partial t}\right]_{t=0} \int \psi(y, T(F))dF + \\
&\qquad + \left[\frac{\partial T(F_t)}{\partial t}\right]_{t=0} \int \psi(y, T(F))d\delta_x = \\
&= \left[\frac{\partial T(F_t)}{\partial t}\right]_{t=0} \int \psi'(y, T(F))dF + \psi(x, T(F));
\end{aligned}$$

da cui, per la definizione di IF (21),

$$\mathrm{IF}(x; T_\psi, F) = \left[\frac{\partial T(F_t)}{\partial t}\right]_{t=0} = \frac{\psi(x, T(F))}{-\int \psi'(y, T(F))F(dy)}. \tag{54}$$

È importante notare che la IF di uno stimatore M è proporzionale alla funzione $\psi(x, .)$ che compare nella definizione dello stesso; è quindi sufficiente che ψ sia limitata (ed il denominatore della IF diverso da zero) affinché lo stimatore relativo T_ψ sia B-robusto.

Se T_ψ è consistente, ossia

$$\int \psi(x, \theta)dF_\theta(x) = 0 \quad \forall\theta, \tag{55}$$

allora si può mostrare che la (54) può essere riscritta

$$\mathrm{IF}(x; T_\psi, F_{\theta_0}) = \frac{\psi(x, T(F))}{\int \psi(y, \theta_0)s(y, \theta_0)F_{\theta_0}(dy)}, \tag{56}$$

dove

$$s(x, \theta_0) := \frac{\partial}{\partial\theta}[\ln f_\theta(x)]_{\theta_0} = \frac{\frac{\partial}{\partial\theta}[f_\theta(x)]_{\theta_0}}{f_{\theta_0}(x)}, \tag{57}$$

è l'usuale *score function* degli stimatori di (massima) verosimiglianza.

Utilizzando la (28), si ottiene la varianza asintotica degli stimatori M:

$$V(T_\psi, F) = \frac{\int \psi^2(x, T(F))F(dx)}{[\int \psi'(y, T(F))F(dy)]^2}, \tag{58}$$

e anche

$$V(T_\psi, F_{\theta_0}) = \frac{\int \psi^2(x, \theta_0)F(dx)}{[\int \psi(y, \theta_0)s(y, \theta_0)F_{\theta_0}(dy)]^2}. \tag{59}$$

9 Stimatori B-robusti ottimi

Nel seguito si farà riferimento alle seguenti ipotesi di lavoro col nome di *condizioni di regolarità*.

Condizioni di regolarità Sia Θ un sottoinsieme aperto e convesso di $\mathbb{R}$, e sia $\{F_\theta; \theta \in \Theta\}$ una famiglia di distribuzioni con densità strettamente positive $f_\theta(x)$ rispetto a qualche misura λ. Sia $\theta_0 \in \Theta$, e si assuma che $s(x, \theta_0)$ esita per ogni $x \in \Omega$, che $\int s(x, \theta_0)dF_{\theta_0}(x) = 0$, e che per l'informazione di Fisher valga $0 < \int s(x, \theta_0)^2 dF_{\theta_0}(x) < \infty$.

Teorema 1 (Hampel [2]**)** *Siano verificate le condizioni di regolarità e sia $b > 0$ una costante reale. Esiste un $a \in \mathbb{R}$ tale che*[7]

$$\tilde{\psi}(x) := [s(x, \theta_0) - a]^b_{-b} \tag{60}$$

soddisfa a $\int \tilde{\psi}(y)dF_0(y) = 0$ ed a $d := \int \tilde{\psi}(y)s(y, \theta_0)dF_0(y) > 0$. Tra le funzioni ψ per le quali valgono

$$\int \psi(y)dF_{\theta_0}(y) = 0, \tag{61}$$

$$\int \psi(y)s(y, \theta_0)dF_{\theta_0}(y) \neq 0, \tag{62}$$

$$\sup_x \left| \frac{\psi(x)}{\int \psi(y)s(y, \theta_0)dF_{\theta_0}(y)} \right| < c := \frac{b}{d} \tag{63}$$

[7]Il significato della notazione che segue nel corpo del testo è

$$[h(x)]^M_m = \begin{cases} M & \text{per } h(x) > M \\ h(x) & \text{per } m \leq h(x) \leq M \\ m & \text{per } h(x) < m \end{cases}.$$

$\tilde{\psi}$ minimizza

$$\frac{\int \psi^2(x)F(dx)}{[\int \psi(y)s(y,\theta_0)F_{\theta_0}(dy)]^2}. \tag{64}$$

Qualunque altra ψ della famiglia specificata che minimizza la (64) è multiplo di $\tilde{\psi}$ quasi ovunque (rispetto a F_{θ_0}).

Osservando le quantità definite nel teorema 1, ci si rende conto delle relazioni che esse hanno con gli stimatori M e con la funzione d'influenza. Infatti, trascurando per il momento l'argomento θ nella funzione ψ che definisce uno stimatore M, si nota che la (61) coincide con la condizione di consistenza di T_ψ, la (62) è una condizione d'esistenza della IF di uno stimatore M (denominatore non nullo), la (63) è la condizione di B-robustezza e (64) è la varianza asintotica di T_ψ. Pertanto la $\tilde{\psi}(.,\theta_0)$ del teorema, tra tutte le funzioni $\psi(.,\theta_0)$ che soddisfano alla condizione di B-robustezza $\gamma^*(\psi, F_{\theta_0}) \leq c(\theta_0) = b(\theta_0)/d(\theta_0)$, è quella che minimizza la varianza asintotica $V(\psi, F\theta_0)$. Si noti che se $b = \infty$, cioè se non s'impone la condizione di robustezza, si ottiene la ψ dello stimatore di messima verosimiglianza. Per passare dal risultato del teorema 1 alla costruzione di uno stimatore M è necessario fissare una funzione $b(\theta)$ e poi trovare la $\tilde{\psi}(.,\theta)$ per ogni θ, che definisca uno stimtore M. Fortunatamente nel caso della stima di parametri di posizione e di scala, è piuttosto semplice trovare lo stimatore M ottimo. Infatti, fissato un valore $b > 0$ costante per ogni θ, è possibile trovare una $\psi(x) = \psi(x,\theta)$ equivariante rispetto a θ.

Nel caso della stima di un parametro di locazione si utilizza una funzione $\psi(x,\theta) = \psi(x-\theta)$, che è sufficiente valutare in $\theta_0 = 0$. La funzione d'influenza del relatico stimatore T_ψ (consistente) è

$$\mathrm{IF}(x; T_\psi, F) = \frac{\psi(x)}{\int \psi(y)s(y,\theta_0)F(dy)}. \tag{65}$$

Lo stimatore M ottimo sarà quindi definito dalla funzione

$$\tilde{\psi}(x) = \left[-\frac{f'(x)}{f(x)} - a\right]_{-b}^{b}. \tag{66}$$

Nel caso in cui la distribuzione F è simmetrica $a = 0$. Se $F = \Phi$ è la normale standard, si ottiene

$$\tilde{\psi}(x) = [x]_{-b}^{b}, \tag{67}$$

che viene spesso chiamata ψ di Huber. Lo stimatore di Huber è quindi quel valore di ϑ per cui vale

$$\sum_{i=1}^{n} [x_i - \vartheta]_{-b}^{b} = 0, \tag{68}$$

che ha però il difetto di non essere invariante a trasformazioni di scala; per renderlo tale si può usare

$$\sum_{i=1}^{n} \left[\frac{x_i - \vartheta}{s} \right]_{-b}^{b} = 0, \tag{69}$$

dove s è una stima (robusta) del fattore di scala. s può essere calcolato per mezzo della MAD (Median Absolute Deviation)

$$\mathrm{MAD}(\{x_i\}_{i=1,\dots,n}) = \mathrm{med}\big(\{|x_i - \mathrm{med}(\{x_i\}_{i=1,\dots,n})|\}_{i=1,\dots,n}\big) \tag{70}$$

magari moltiplicata per 1.483, per renderla uno stimatore corretto della deviazione standard della normale.

Osservazioni

Quando F e ψ sono simmetriche, per gli stimatori M di posizione si dimostra [3][10] che se ψ è strettamente monotòna valogono le seguenti proprietà: (i) l'equazione che definisce lo stimatore M ha un'unica soluzione, (ii) se ψ è limitata lo stimatore è B-robusto, qualitativamente robusto, e con punto di rottura $\delta^* = 0.5$, (iii) se ψ non è limitata lo stimatore non è né B-robusto, né qualitativamente robusto, e ha punto di rottura $\delta^* = 0$.

Se la soluzione dell'equazione che definisce lo stimatore M non è unica è consigliabile prendere la soluzione più vicina alla mediana (o dare la mediana come valore iniziale dell'algoritmo di ottimizzazione).

Sotto l'ipotesi $F = \Phi$, lo stimatore di Huber è B-robusto (ottimo), qualitativamente robusto, ed il punto di rottura è $\delta^* = 0.5$.

Quando $b \downarrow 0$, lo stimatore di Huber tende alla mediana, che è lo stimatore M più B-robusto.

Riferimenti bibliografici

[1] Fernhotz, L. T. (1983). *Von Mises Calculus for Statistical Functionals*. Lecture Notes in Statistics 19. Sprnger, New York.

[2] Hampel, F. R. (1968). Contributions to the theory of robust estimation. Ph.D. thesis. University of California, Berkeley.

[3] Hampel, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Stat* 42, pp. 1887–1896.

[4] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Assoc.* 69, pp. 383–393.

[5] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1986). *Robust Statistics. The Approach Based on Influence Functions.* John Wiley & Sons, New York.

[6] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* 35, pp. 73–101.

[7] Huber, P. J. (1965). A robust version of the probability ratio test. *Ann. Math. Stat.* 36, pp. 1753–1758.

[8] Huber, P. J. (1972). Robust statistics: A review. *Ann. Math. Stat.* 43, pp. 1041–1067.

[9] Huber, P. J. (1977). *Robust Statistical Procedures*. Society for Industrial and Applied Mathematics, Philadelphia.

[10] Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

[11] Rey, W. J. J. (1983). *Introduction to Robust and Quasi-Robust Statistical Methods*. Springer-Verlag, Berlin.

[12] Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics*, I. Olkin (ed.). Stanford University Press, Stanford, pp. 448–485.

[13] Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison–Wesley, Reading, Mass.

[14] von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Stat.* 18, 309–348.