METODI DI RICAMPIONAMENTO

INTRODUZIONE

Sono una delle più interessanti applicazioni inferenziali delle simulazioni stocastiche e della generazione di numeri casuali

I metodi di ricampionamento si sono diffusi a partire dagli anni ’60, derivati concettualmente dai metodi “Monte Carlo”

Lo sviluppo dei metodi “Monte Carlo” è avvenuto principalmente a partire dagli anni ‘80 del secolo scorso, a seguito dei progressi delle tecnologie informatiche e dell’aumento della potenza dei computer

La loro utilità è legata allo sviluppo di metodi non parametrici, in situazioni in cui i metodi dell’inferenza classica non possono essere correttamente applicati

INTRODUZIONE

Questi metodi vanno sotto l'etichetta "*Computational Intensive*"

- Sono metodi che ripetono semplici operazioni un numero elevato di volte
- Generano numeri casuali da assegnare a variabili casuali o a campioni casuali
- Richiedono maggior "*Tempo Macchina*" al crescere delle operazioni ripetute
- Sono molto semplici da implementare ed una volta implementati sono automatici

METODI PARAMETRICI E NON PARAMETRICI

L'inferenza classica si è sviluppata in ambito parametrico

Una metodologia inferenziale è detta parametrica qualora sia basata sull'assunto che la v.c. X da cui proviene il campione abbia legge di probabilità $F(x; \theta)$, tipicamente *Normale*, nota a meno del valore dei parametri θ

VANTAGGI	SVANTAGGI
Bagaglio teorico	Non sono utilizzabili quando non si hanno informazioni su $F(x)$
Utilizzo di tavole note	Non è sempre possibile ricorrere alle approssimazioni asintotiche
Costo computazionale basso	Non sono sempre soddisfatte le assunzioni di omoschedasticità, normalità, i.i.d., ecc.

METODI PARAMETRICI E NON PARAMETRICI

METODI NON PARAMETRICI

Mancanza di informazioni a priori su *F(x)*

Riutilizzo iterato del campione

METODI PARAMETRICI E NON PARAMETRICI

Produzione e impiego di versioni casualizzate del campione originario (le sole informazioni note) ottenute ri-campionando le unità

VANTAGGI	SVANTAGGI
Sono più "potenti" quando le assunzioni non sono verificate	Non sono più "precisi" dei metodi classici quando le assunzioni sono vere
Permettono di valutare e/o migliorare l'accuratezza di uno stimatore	Forniscono soluzioni approssimate poiché basate su simulazioni

"*torturare i dati al fine di farli confessare*"

METODI DI RICAMPIONAMENTO

METODI DI SIMULAZIONE

Un processo di simulazione consiste nel:

1. Costruire un modello che sia in grado di imitare il sistema in esame
2. Generare dal modello numerosi campioni possibili e studiarne il comportamento
3. Analizzare i risultati evidenziando le decisioni alternative

La simulazione consente non solo di seguire l'evolversi di un processo ma anche di prevedere situazioni future

Per poter implementare una simulazione bisogna saper simulare tramite computer l'estrazione di numeri casuali provenienti da particolari distribuzioni, ossia determinazioni (indipendenti) di variabili casuali

METODI DI RICAMPIONAMENTO

JACKKNIFE

Jackknife = coltello a serramanico (coltellino svizzero)

1949: M. H. Quenouille

1958: Tukey

OBIETTIVO	ridurre le distorsioni sistematiche (che dipendono dai dati campionari) nella stima delle statistiche di una popolazione, fornendone l'errore standard

METODI DI RICAMPIONAMENTO

JACKKNIFE

$$x = (x_1, x_2, ..., x_n) \qquad X \sim F(x;\theta)$$

***F* è ignota!!!**

Non essendo nota la distribuzione della variabile non è nota neanche la distribuzione dello stimatore

$$\hat{\Theta} \sim F(x;\hat{\Theta})$$

METODI DI RICAMPIONAMENTO

JACKKNIFE

I campioni *Jackknife* sono costruiti lasciando fuori ogni volta dal campione originario un'osservazione x_i:

$$x_{(i)} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

Quindi si ottengono *n* campioni di dimensione *m=n-1*

Esempio

Consideriamo un campione di numerosità *n=5*, che produce 5 campioni *Jackknife* di numerosità *m=4*

$$x_{(1)} = (x_2, x_3, x_4, x_5)$$
$$x_{(2)} = (x_1, x_3, x_4, x_5)$$
$$x_{(3)} = (x_1, x_2, x_4, x_5)$$
$$x_{(4)} = (x_1, x_2, x_3, x_5)$$
$$x_{(5)} = (x_1, x_2, x_3, x_4)$$

METODI DI RICAMPIONAMENTO

JACKKNIFE

Sul generico *i-esimo* campione *Jackknife* si ricalcola lo pseudo-valore $\hat{\theta}_{(i)}$, identico alla stima $\hat{\theta}$

La procedura viene iterata *n* volte su ciascuno dei campioni *Jackknife* disponibili

$$x=(x_1,x_2,\quad,x_n)$$

$$x_{(1)}=(x_2,\quad,x_n) \rightarrow \hat{\theta}_{(1)}$$

$$x_{(i)}=(x_1,\quad,x_{i-1},x_{i+1},\quad,x_n) \rightarrow \hat{\theta}_{(i)}$$

$$x_{(n)}=(x_1,\quad,x_{n-1}) \rightarrow \hat{\theta}_{(n)}$$

METODI DI RICAMPIONAMENTO

JACKKNIFE

$\hat{\Theta} = f(X_1, \quad , X_n)$ Stimatore

$\hat{\theta} = f(x_1, \quad , x_n)$ Stima basata sul campione originario

$\hat{\theta}_{(i)} = f(x_1, \quad , x_{i-1}, x_{i+1}, \quad , x_n)$ Pseudo-valori *Jackknife*

$i = 1, \quad , n$

$\hat{\theta}_{(\bullet)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(i)}$ Stima *Jackknife*

È una simulazione di:

$E(\hat{\Theta})$

basata sul ricampionamento

METODI DI RICAMPIONAMENTO

JACKKNIFE Come si corregge la distorsione?

Con l'obiettivo di valutare, e possibilmente ridurre, la distorsione dello stimatore $\hat{\Theta}$, si costruisce la stima *Jackknife* della distorsione

B stima per la distorsione basata sul ricampionamento

$$\hat{B}_J = (n-1)\left(\hat{\theta}_{(\bullet)} - \hat{\theta}\right)$$

Gioca il ruolo di peso tenendo conto che il ricampionamento prevede ampiezze campionarie inferiori rispetto al campione originario

$$\hat{\theta}_{J-corr} = \hat{\theta} - \hat{B}_J\left(\hat{\Theta}\right) = \hat{\theta} - (n-1)\left(\hat{\theta}_{(\bullet)} - \hat{\theta}\right) = n\hat{\theta} - (n-1)\hat{\theta}_{(\bullet)}$$

Stima *Jackknife* corretta per θ

METODI DI RICAMPIONAMENTO

JACKKNIFE

Esempio 1

$$\theta = \mu = E(X)$$

Parametro oggetto di stima: **media**

$$\hat{\theta} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Stima: media campionaria

Il generico pseudo-valore è pari a:

$$\hat{\theta}_{(i)} = \frac{1}{n-1}\sum_{j\neq i=1}^{n} x_j = \frac{1}{n-1}\left(\sum_{j=1}^{n} x_j - x_i\right) = \frac{n\bar{x}}{n-1} - \frac{x_i}{n-1}$$

METODI DI RICAMPIONAMENTO

JACKKNIFE

La stima *Jackknife* è pari a:

$$\hat{\theta}_{(\bullet)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{n\bar{x}}{n-1} - \frac{x_i}{n-1}\right) = \frac{n\bar{x}}{n-1} - \frac{\bar{x}}{n-1} = \frac{n\bar{x}-\bar{x}}{n-1} = \frac{(n-1)\bar{x}}{n-1} = \bar{x}$$

La stima *Jackknife* della distorsione dello stimatore media campionaria risulta nulla

$$\hat{B}_J = (n-1)\left(\hat{\theta}_{(\bullet)} - \hat{\theta}\right) = (n-1)(\bar{x}-\bar{x}) = 0$$

Quindi nel caso della media campionaria:

$$\hat{\theta}_{J-corr} = \hat{\theta} - \hat{B}_J\left(\hat{\Theta}\right) = \bar{x} - 0 = \bar{x}$$

METODI DI RICAMPIONAMENTO

JACKKNIFE

Esempio 2

$$\theta = \sigma^2 = VAR(X)$$

Parametro oggetto di stima: **varianza**

$$\hat{\theta} = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Stima: varianza campionaria

Si dimostra che la stima *Jackknife* è pari a:

$$\hat{\theta}_{(\bullet)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} = \frac{n(n-2)}{(n-1)^2}s^2$$

La stima *Jackknife* della distorsione dello stimatore varianza campionaria risulta:

$$\hat{B}_J = (n-1)\left(\hat{\theta}_{(\bullet)} - \hat{\theta}\right) = -\frac{s^2}{n}$$

Quindi nel caso della varianza campionaria:

$$\hat{\theta}_{J-corr} = \hat{\theta} - \hat{B}_J\left(\hat{\Theta}\right) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

METODI DI RICAMPIONAMENTO

JACKKNIFE

La stima *Jackknife* dell'errore standard è:

$$\hat{\sigma}_J = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\bullet)}\right)^2}$$

Tale quantità sarà utilizzata per la costruzione di intervalli di fiducia per i parametri

$$\hat{\theta} \pm z_\alpha \hat{\sigma}_J$$

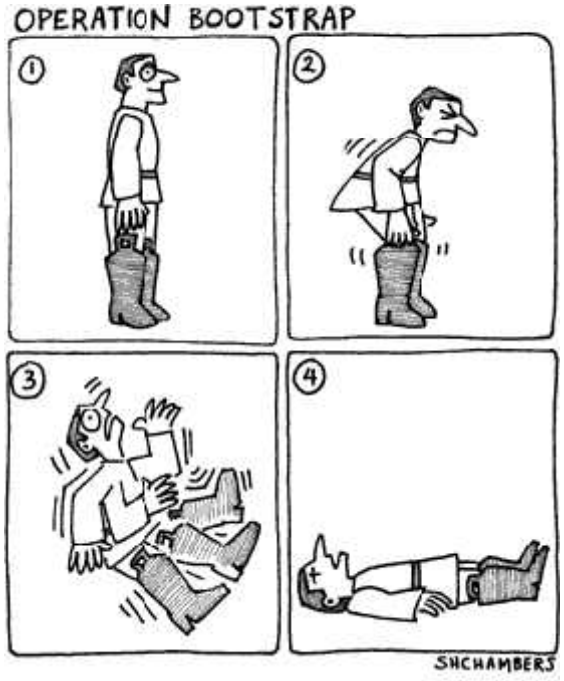
$$\hat{\theta} \pm t_\alpha \hat{\sigma}_J$$

METODI DI RICAMPIONAMENTO

BOOTSTRAP

Bootstrap = stringhe da stivali

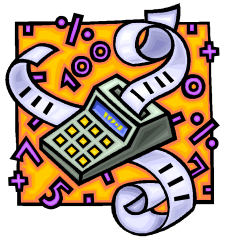
1979: B. Efron



"*Adventures of Baron Munchausen*"
(R. E. Raspe)

"*to pull oneself up by one's Bootstrap*"

È un metodo *Computer Intensive*



Nasce come evoluzione del metodo *Jackknife*, per stimare l'Errore Standard dello stimatore di un parametro

METODI DI RICAMPIONAMENTO

BOOTSTRAP

Si basa sul fatto paradossale che l'unico campione disponibile serve per generarne molti altri e per costruire la distribuzione teorica di riferimento

Usa i dati del campione originario per calcolare una statistica e stimare la sua distribuzione campionaria senza fare alcuna ipotesi sul modello della distribuzione

JACKKNIFE

METODI DI RICAMPIONAMENTO

BOOTSTRAP

Per poter conoscere la distribuzione si utilizza il principio *Plug-in* (Principio di Sostituzione)

La stima per θ si costruisce sostituendo alla funzione di ripartizione incognita F della popolazione il suo equivalente empirico:

$$\hat{F}$$

Funzione di ripartizione del campione ottenuta costruendo una distribuzione di frequenze di tutti i valori che esso può assumere in quella situazione sperimentale

METODI DI RICAMPIONAMENTO

BOOTSTRAP

I campioni *Bootstrap* sono costruiti estraendo con ripetizione *n* unità dal campione originario di numerosità *n*

$$x^*_{(b)} = \left(x^*_1, \quad , x^*_n\right)$$

È quindi possibile estrarre un numero β di campioni *Bootstrap*, diversi fra loro per almeno un elemento, con:

$$\beta = \binom{2n-1}{n}$$

Per *n=5* sono *126*

Per *n=6* sono *462*

Per *n=10* sono *92378*

Non potendo tener conto di tutti i β possibili campioni *Bootstrap*, si terrà conto solo di una parte (*B*) di questi

METODI DI RICAMPIONAMENTO

BOOTSTRAP

Esempio

Consideriamo un campione di numerosità *n=5*, che può produrre fino a *126* campioni *Bootstrap* di numerosità *n=5*

$$x^*_{(1)} = (x_1, x_2, x_3, x_4, x_4)$$

$$x^*_{(2)} = (x_1, x_3, x_3, x_4, x_5)$$

$$x^*_{(b)} = (x_1, x_2, x_4, x_5, x_5)$$

$$x^*_{(B)} = (x_1, x_1, x_2, x_5, x_5)$$

Ogni campione *Bootstrap* permette di ottenere una stima della statistica desiderata:

$$\hat{\theta}^*_{(b)} = f(x^*_1, \quad , x^*_n) \qquad \hat{\theta}^*_{(\bullet)} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^*_{(b)}$$

METODI DI RICAMPIONAMENTO

BOOTSTRAP

La distribuzione di una statistica raccoglie i suoi valori da molti ricampionamenti

Il campione originario rappresenta la popolazione, mentre i suoi ricampionamenti rappresentano ciò che otterremmo se prendessimo molti campioni dalla popolazione

$$x=(x_1,x_2,\quad ,x_n)$$

$$x^*_{(1)}=\left(x_1^{*1},\quad ,x_n^{*1}\right) \Rightarrow \hat{\theta}^*_{(1)}$$

$$x^*_{(b)}=\left(x_1^{*b},\quad ,x_n^{*b}\right) \Rightarrow \hat{\theta}^*_{(b)}$$

$$x^*_{(B)}=\left(x_1^{*B},\quad ,x_n^{*B}\right) \Rightarrow \hat{\theta}^*_{(B)}$$

METODI DI RICAMPIONAMENTO

BOOTSTRAP

Il *Bootstrap* ci permette di studiare:

- La forma di una distribuzione (*shape*)
- Il valore centrale di una distribuzione (*center*)
- La distorsione dello stimatore (*bias*)
- La stima dell'errore standard (*spread*)

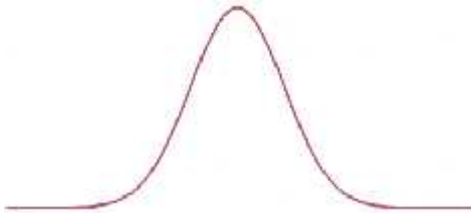
METODI DI RICAMPIONAMENTO

BOOTSTRAP

Shape (forma)

Il Teorema del Limite Centrale dice che la distribuzione campionaria di una media campionaria è approssimativamente Normale per *n* abbastanza grande

Così la forma della distribuzione *Bootstrap* è vicina alla forma che ci attendiamo dalla distribuzione campionaria



Poiché la forma della distribuzione *Bootstrap* si approssima a quella della distribuzione campionaria, è possibile utilizzare la distribuzione *Bootstrap* per verificare la Normalità della distribuzione campionaria

METODI DI RICAMPIONAMENTO

BOOTSTRAP

Il *Bootstrap* ci permette di studiare:

- La forma di una distribuzione (*shape*)
- Il valore centrale di una distribuzione (*center*)
- La distorsione dello stimatore (*bias*)
- La stima dell'errore standard (*spread*)

METODI DI RICAMPIONAMENTO

BOOTSTRAP

Center (centro)

Il centro della distribuzione *Bootstrap* non è lo stesso del centro della distribuzione campionaria

La distribuzione campionaria di una statistica usata per stimare il parametro è centrata al valore del parametro stesso (valore incognito della popolazione)

La distribuzione *Bootstrap* è centrata al valore della statistica per il campione originario

METODI DI RICAMPIONAMENTO

BOOTSTRAP

Il *Bootstrap* ci permette di studiare:

- La forma di una distribuzione (*shape*)
- Il valore centrale di una distribuzione (*center*)
- La distorsione dello stimatore (*bias*)
- La stima dell'errore standard (*spread*)

METODI DI RICAMPIONAMENTO

BOOTSTRAP

Bias (distorsione)

La stima *Bootstrap* della distorsione è data dalla differenza tra il valore medio calcolato con le *B* estrazioni del metodo *Bootstrap* e quello calcolato con gli *n* dati originari

$$\hat{B}_B = \hat{\theta}^*_{(\bullet)} - \hat{\theta} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^*_{(b)} - \hat{\theta}\right)$$

- È il valore calcolato sul campione osservato
- È la media dei valori calcolati sulle *B* replicazioni *Bootstrap*

METODI DI RICAMPIONAMENTO

BOOTSTRAP

Il *Bootstrap* ci permette di studiare:

- La forma di una distribuzione (*shape*)
- Il valore centrale di una distribuzione (*center*)
- La distorsione dello stimatore (*bias*)
- La stima dell'errore standard (*spread*)

METODI DI RICAMPIONAMENTO

BOOTSTRAP

Spread (errore standard)

La Deviazione Standard di *B* replicazioni *Bootstrap* è lo stimatore dell'Errore Standard

$$\hat{\sigma}_B = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^*_{(b)} - \hat{\theta}^*_{(\bullet)}\right)^2}$$

Tale quantità sarà utilizzata per la costruzione di intervalli di fiducia per i parametri

$$\hat{\theta} \pm z_\alpha \hat{\sigma}_B$$

$$\hat{\theta} \pm t_\alpha \hat{\sigma}_B$$

Il calcolo di intervalli di fiducia con i percentili ha il limite di creare intervalli non simmetrici

METODI DI RICAMPIONAMENTO

BOOTSTRAP

Quanti campioni *Bootstrap* bisogna prendere?

Empiricamente:

$25 < B < 200$	Serve ad avere delle informazioni iniziali
$B = 200$	Serve a stimare l'Errore Standard
$B = 500$	Serve a creare un intervallo di fiducia

Lo stimatore *Bootstrap* dell'Errore Standard è consistente e non distorto rispetto allo stimatore ideale (calcolato considerando tutti i possibili campioni *Bootstrap*)

$$\hat{\sigma}_B\left(\hat{\Theta}\right) \xrightarrow{p} \hat{\sigma}_\beta\left(\hat{\Theta}\right)$$

METODI DI RICAMPIONAMENTO

CONFRONTO FRA BOOTSTRAP E JACKKNIFE

L'obiettivo di entrambi i metodi è quello di misurare l'accuratezza (Errore Standard, Distorsione) per una statistica $\hat{\Theta}$ da un set di dati

$$x = (x_1, x_2, \quad , x_n)$$

JACKKNIFE

Il metodo *Jackknife resampling* genera campioni di dimensione inferiore ad *n* senza *re-immissione*

BOOTSTRAP

Il metodo *Bootstrap resampling* genera campioni di dimensione pari ad *n* con *re-immissione*

METODI DI RICAMPIONAMENTO

CONFRONTO FRA BOOTSTRAP E JACKKNIFE

Stima dell'Errore Standard

$$\hat{\sigma}_J = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\bullet)}\right)^2} \qquad \hat{\sigma}_B = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^*_{(b)} - \hat{\theta}^*_{(\bullet)}\right)^2}$$

La stima *Jackknife* rappresenta la devianza degli *n* valori dei campioni, a meno di un fattore di scala (*inflaction factor*) $\frac{n-1}{n}$

$$\frac{n-1}{n} > \frac{1}{B-1}$$

Intuitivamente tale fattore è necessario in quanto la deviazione del *Jackknife* tende ad essere molto più piccola di quella del *Bootstrap* (ogni campione di *Jackknife* è molto più simile al campione originario)

METODI DI RICAMPIONAMENTO

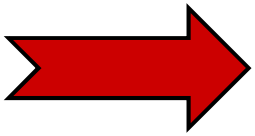
CONFRONTO FRA BOOTSTRAP E JACKKNIFE

Stima della distorsione

$$\hat{B}_J = (n-1)\left(\hat{\theta}_{(\bullet)} - \hat{\theta}\right)$$

$$\hat{B}_B = \hat{\theta}^*_{(\bullet)} - \hat{\theta} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^*_{(b)} - \hat{\theta}\right)$$

Le stime della distorsione sono strettamente legate fra loro



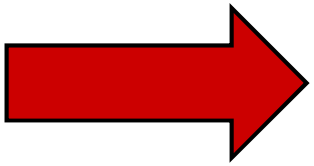
Teorema di Efron (1982)

$$\hat{B}_J = \frac{n}{n-1}\left(\hat{B}_B\right)$$

METODI DI RICAMPIONAMENTO

CONFRONTO FRA BOOTSTRAP E JACKKNIFE

Il metodo *Jackknife* usa meno informazioni del *Bootstrap*



Il *Jackknife* è un'approsimazione del *Bootstrap*

Il valore ottenuto da *n* replicazioni *Jackknife* coincide con il valore di *Bootstrap* a meno del fattore di scala $\sqrt{\frac{n-1}{n}}$

METODI DI RICAMPIONAMENTO

BIBLIOGRAFIA - SITOGRAFIA

[1] A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation – B.Efron, G.Gong – The American Statistician, Vol. 37, N° 1, Feb. 1983, pp. 36-48

[2] Bootstrap Methods and Permutation Tests – T.Hesterberg, D.S.Moore, S. Monaghan, A.Clipson, R.Epstein - 2005

[3] Nonparametric Estimates of Standard Error: the Jackknife, the Bootstrap and Other Metohds – B.Efron – Biometrika, 1981, 68, 3, pp. 589-99

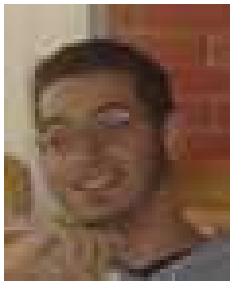
[4] Resampling Methods: Randomization Tests, Jackknife and Bootstrap Estimators - B.Walsh - Lecture Notes for EEB 596z, 2000

[S1] http://mason.gmu.edu/~csutton/gong789topi.html

[S2] http://pareonline.net/getvn.asp?v=8&n=19

METODI DI RICAMPIONAMENTO

Grazie a:



Enrico Infante

Nicolina Lippiello