Laboratorio di Statistica Computazionale - 3

Matteo Grigoletto[1]

[1] Il materiale presentato è basato sugli appunti di Domenico De Stefano.

Argomenti di oggi

- Bootstrap
  - non parametrico
  - semi-parametrico
  - parametrico
- Applicazioni

Alcuni esercizi per oggi /1

- Esercizio n. 1. L'idea di questo esercizio è studiare la distorsione e la varianza di uno stimatore $\hat{\theta}$ di un parametro θ, attraverso la distorsione e la varianza di $\hat{\theta}^\star$ rispetto a $\hat{\theta}$.
  - **a.** Per l'analisi, si generino 50 osservazioni da v.c. i.i.d. con distribuzione $\mathcal{N}(0, 1)$. Si finga poi di non sapere da che distribuzione provengono i dati, né quanto valgono la distorsione e la varianza della media campionaria e le si stimi con il metodo bootstrap non parametrico.
  - **b.** Si aumenti la varianza della distribuzione normale da cui provengono le osservazioni. Cosa succede alle stime bootstrap della distorsione e della varianza?
  - **c.** Lasciando invariato tutto il resto, si aumenti il numero di campioni bootstrap estratti. Cosa succede alle stime bootstrap della distorsione e della varianza?

Alcuni esercizi per oggi /2

- Esercizio n. 2. Facendo riferimento agli stessi dati dell'esercizio n. 1, si supponga ora che il parametro di interesse θ sia la mediana. Ci chiediamo quanto è accurato lo stimatore $\hat{\theta}$ dato dalla mediana campionaria.
  - **a.** Si provi a risolvere il problema senza ricorrere a tecniche di simulazione ed usando risultati asintotici, supponendo ancora di non conoscere la distribuzione da cui provengono i dati[†].
  - **b.** Si utilizzi ora, allo stesso scopo, il metodo bootstrap non parametrico.

[†] Per n v.c. i.i.d. con densità f avente vera mediana $\tilde{\mu}$, sotto alcune assunzioni, la distribuzione asintotica della mediana campionaria è $\mathcal{N}\left(\tilde{\mu}, 1/\{4\,n\,[f(\tilde{\mu})]^2\}\right)$. La funzione **sm.density** nella biblioteca **sm** calcola stime non parametriche della funzione di densità. In particolare, il comando "`sm.density(dati, eval.points=med)$estimate`" restituisce il valore della densità, stimata sulla base delle osservazioni in "`dati`", calcolata nel punto "`med`".

Alcuni esercizi per oggi /3

- Esercizio n. 3.
  - **a.** Il dataset "faithful" (richiamabile in **R** con l'istruzione **data(faithful)**) contiene le durate delle eruzioni di un geyser chiamato *Old Faithful* ed i tempi di attesa fra un'eruzione e la successiva. Si usi il bootstrap non parametrico per costruire un intervallo di confidenza di livello 0.95 per la durata mediana delle eruzioni.
  - **b.** Supponiamo di essere incerti se usare $B = 500$, $B = 1000$, $B = 5000$ oppure $B = 10000$ per il numero di replicazioni bootstrap. Il valore di B viene considerato sufficientemente elevato se, riapplicando più volte la procedura bootstrap, i risultati differiscono tra loro meno di una soglia valutata accettabile. Nel caso presente, supponiamo di usare il criterio "il valore di B è accettabile se, ripetendo la procedura più volte, si ottengono intervalli di confidenza i cui estremi hanno deviazione standard inferiore a 0.005". Si implementi una procedura che ci permette di scegliere B sulla base di tale criterio.

Alcuni esercizi per oggi /4

- Esercizio n. 4. Per gli stessi dati dell'esercizio precedente, si usi il bootstrap non parametrico per stimare $P(|\bar{Y} - \mu| > 0.2)$, dove $\bar{Y}$ è la media campionaria e μ la media della popolazione. In altri termini, vogliamo sapere qual è la probabilità di commettere un errore assoluto superiore a 0.2 quando stimiamo la media.

Alcuni esercizi per oggi /5

- Esercizio n. 5. Uno dei parametri che regola il funzionamento globale di un impianto chimico è la temperatura dell'acqua di raffreddamento. Si vuole capire se la perdita di produzione (in percentuale) dipende da tale temperatura in modo lineare. Dato che ci si attende che alcuni valori si discostino notevolmente dalla retta di regressione (fermate parziali dell'impianto) si adatta un modello di regressione robusta[‡]. Si usi il bootstrap non parametrico, semi-parametrico e parametrico, per stimare la distorsione degli stimatori dei due coefficienti di regressione e si propongano infine due stime corrette. I dati sono contenuti nel file `impianto.dat`

[‡]Un metodo alternativo di regressione detto LTS (*Least Trimmed Squares*) si basa sulla minimizzazione di una porzione $q < n$ dei quadrati dei residui ordinati in ordine crescente. Una tecnica di tale genere sopporta quindi un certo numero di osservazioni molto distanti dalla media senza fornire stime distorte dei parametri (ossia, tale metodo è robusto rispetto a tale tipo di osservazioni). Si parla in questo caso di regressione robusta. In **R** la funzione che esegue una regressione robusta è **ltsreg** disponibile all'interno della biblioteca **MASS**. La sintassi è analoga a **lm**. La scelta ottimale di q, di cui noi non ci occupiamo, rientra nel solito dilemma tra efficienza e distorsione.

Alcuni esercizi per oggi /6

- Esercizio n. 6. Si supponga che i dati nel file `esponenziale.dat` provengano dalle v.c. i.i.d. $Y_1, \ldots, Y_{300}$, aventi distribuzione esponenziale di media $1/\lambda$.
  - **a.** Si fornisca, usando il bootstrap parametrico, un intervallo di confidenza di livello 0.95 per λ.
  - **b.** Si faccia la stessa cosa del punto precedente, implementando una propria funzione per il calcolo dei quantili.

Note esercizio 5

- In questo esercizio, l'approccio parametrico merita particolare attenzione. In questo approccio, dobbiamo generare i dati con X fissata pari ai valori x osservati per la temperatura e

$$Y^\star = \hat{\beta}_0 + \hat{\beta}_1 \cdot x + \epsilon^\star \ ,$$

dove $\hat{\beta}_0$ e $\hat{\beta}_1$ sono le stime ottenute con la regressione robusta. Per $\epsilon^\star$ non ha senso assumere una distribuzione normale, che non riuscirebbe a rappresentare i valori anomali e sarebbe quindi del tutto irrealistica. Poiché, quando si presenta l'anomalia, i valori di ϵ sono negativi (fermata parziale dell'impianto), potremmo porre

$$\epsilon^\star = (1 - V) \cdot \epsilon_1^\star - V \cdot \epsilon_2^\star \ ,$$

dove $V \sim \text{Bin}(1, \hat{\gamma})$, $\epsilon_1^\star \sim \mathcal{N}(0, \hat{\sigma}^2)$ e $\epsilon_2^\star \sim \text{Exp}(\hat{\lambda})$. Qui, γ è la probabilità di avere un valore anomalo, σ^2 è la varianza dei dati "ordinari" e λ caratterizza la distribuzione dei dati anomali.