Laboratorio di Statistica Computazionale - 4

Matteo Grigoletto

Argomenti di oggi

- Bootstrap (di nuovo).
  - Test.
  - Bootstrap lisciato.

Alcuni esercizi per oggi /1

- Esercizio n. 1. Per verificare l'efficacia di due farmaci, 43 pazienti con asma cronica ma in forma lieve sono stati suddivisi casualmente in due gruppi: al primo, di 26 pazienti, è stato somministrato il primo trattamento; al secondo, di 17 pazienti, è stato somministrato il secondo. Ad ogni paziente è stato misurato il picco di flusso espiratorio (la massima velocità con cui l'aria può essere espulsa dai polmoni dopo una inspirazione profonda) sia all'inizio dello studio che dopo 30 giorni. Il file `asma.dat` contiene per ogni paziente il trattamento ricevuto (prima colonna, variabile `trattamento` con valori '1' o '2') e la differenza tra il picco di flusso misurato dopo 30 giorni e quello iniziale (seconda colonna, variabile risposta). Si tenga presente che un picco di flusso più alto indica che il trattamento è efficace. L'interesse dei medici consiste nel capire se c'è una differenza significativa in posizione tra i due gruppi.
  - **a.** Verificare, utilizzando l'usuale test t a due campioni, l'ipotesi che le medie dei gruppi siano le stesse (si verifichi l'assunzione di normalità).
  - **b.** Verificare la stessa ipotesi utilizzando un approccio bootstrap nonparametrico.
  - **c.** Verificare, sempre utilizzando un approccio bootstrap nonparametrico, l'ipotesi che le vere mediane dei picchi di flusso siano uguali nei due gruppi.
  - **d.** Commentare i risultati.

Alcuni esercizi per oggi /2

- Esercizio n. 2. Si supponga che $X_1, \ldots, X_{n_x}$ siano v.c. i.i.d. con distribuzione Gamma(α_x, λ_x) e $Y_1, \ldots, Y_{n_x}$ siano v.c. i.i.d. con distribuzione Gamma(α_y, λ_y). Le v.c. X sono indipendenti dalle Y. Nel seguito, poniamo $\mu_x = E(X)$ e $\mu_Y = E(Y)$.
  - **a.** Si valuti se applicare il test t di Student (con la correzione di Welch) per verificare $H_0 : \mu_x = \mu_Y$ contro $H_1 : \mu_x \neq \mu_Y$, nella condizione descritta, porta ad un test con livello di significatività pari a quello nominale[1].
  - **b.** Si studi ancora il livello di significatività nominale esattamente come nel punto precedente, ma usando un test bootstrap.
  - **c.** Si ripetano i passi a. e b., ma con l'obiettivo di studiare le potenze dei due test[2].
  - **d.** Si commentino i risultati.

[1] In particolare, si usino $n_x = n_y = 15$ ed un livello di significatività nominale α pari a 0.05. Inoltre, si fissino $\mu_x = \mu_y = 3$ e $\alpha_x = \alpha_y = 2$ (questo implica $\lambda_x = \lambda_y = 2/3$). La verifica deve essere basata su $M = 400$ iterazioni Monte Carlo.

[2] Si usino $\mu_x = 3$ e $\mu_y = 5$.

Alcuni esercizi per oggi /3

- Esercizio n. 3. Nel 1882 Simon Newcomb fece un esperimento per misurare la velocità della luce. I valori sottostanti rappresentano le misurazioni ottenute.

  28 -44 29 30 26 27 22 23 33 16 24 29 24 40 21 31 34 -2 25 19

  Nell'unità di misura usata, la velocità della luce attualmente nota è 33.02. I dati di Newcomb sono coerenti con questo valore?

  a. Chiamiamo Y la v.c. che genera i dati di Newcomb e poniamo $\mu = E(Y)$. Si applichi un test t di Student per verificare $H_0 : \mu = 33.02$, verificando però l'assunzione di normalità.

  b. Si applichi un test bootstrap, con $B = 100000$, per verificare la stessa ipotesi.

  c. Nei dati di Newcomb sembrano essere presenti alcuni valori anomali. Si costruisca un test bootstrap (sempre con $B = 100000$) per verificare l'ipotesi $H_0 : \text{Mediana}(Y) = 33.02$.

Alcuni esercizi per oggi /4

- Esercizio n. 4.
  - **a.** Usando gli stessi dati dell'esercizio n. 3, si risponda alle domande a., b. e c. usando degli opportuni intervalli di confidenza, con grado di fiducia 0.99.
  - **b.** Si applichi un test bootstrap per verificare $H_0 : \mu = 33.02$, usando però la funzione test non studentizzata $T = \bar{Y} - 33.02$.
  - **c.** Si commentino i risultati, confrontandoli con quelli ottenuti nell'esercizio n. 3.

Alcuni esercizi per oggi /5

- Esercizio n. 5. Il file `houseprices.dat` contiene le superfici (X, in piedi quadrati) ed in prezzi (Y, in migliaia di dollari australiani) di un campione di case nei pressi di Canberra, Australia.

  Si consideri il modello lineare semplice $Y = \beta_1 + \beta_2 X + \varepsilon$. Si desidera saggiare l'ipotesi che l'aumento di un piede quadrato della superficie procuri, in media, un aumento del prezzo della casa pari a 400 dollari[3].

  a. Si verifichi l'ipotesi con l'usuale intervallo di confidenza per β_2, verificando però l'assunzione di normalità per ε.
  b. Si verifichi la stessa ipotesi, costruendo intervalli di confidenza per β_2, basati sul bootstrap nonparametrico, parametrico e semiparametrico[4].
  c. Si commentino i risultati.

[3] Quindi, l'ipotesi è $H_0 : \beta_2 = 0.4$.

[4] Si noti che, in questo caso, non è banale capire come costruire un test bootstrap per β_2. Infatti, come imponiamo H_0 nel mondo bootstrap, in modo da calcolare la distribuzione di una funzione test sotto tale ipotesi? Dobbiamo certamente escludere l'approccio non parametrico. Per quanto riguarda l'approccio semiparametrico, potremmo ad esempio pensare di usare $Y^\star = \hat{\beta}_1 + e^*$, dove $\hat{\beta}_1$ e la distribuzione di e^* sono stimate, nel mondo reale, imponendo $\beta_2 = 0.4$. Questa imposizione, tuttavia, cambia la distribuzione degli pseudo-dati in maniera non banale (ossia anche riguardo ad aspetti che non sono necessariamente legati al fatto che $\beta_2 = 0.4$). Per il bootstrap parametrico il problema è analogo. Questo ci spinge, per semplicità, ad usare gli intervalli di confidenza, invece dei test, per risolvere il problema corrente.

Alcuni esercizi per oggi /6

- Esercizio n. 6. Siano $Y_1, \ldots, Y_n$ v.c. i.i.d. con distribuzione $\mathcal{N}(0,1)$. Lo stimatore $\hat{q}_\alpha$[5] del quantile q_α è distorto. Si vuole valutare la capacità delle tecniche bootstrap di correggere tale distorsione. Si fissino $n = 100$ e $\alpha = 0.3$. Inoltre, si usino $B = 1000$ replicazioni bootstrap.
  - **a.** Si generino, via Monte Carlo, n osservazioni da un distribuzione $\mathcal{N}(0,1)$. Si usi poi il bootstrap non parametrico discreto (basato sulla FdR empirica) per stimare la distorsione $\delta = E(\hat{q}_\alpha) - q_\alpha$ ed ottenere quindi lo stimatore $\hat{q}'_\alpha = \hat{q}_\alpha - \hat{\delta}$.
  - **b.** Si usi il bootstrap non parametrico lisciato per raggiungere lo stesso obiettivo del punto a.
  - **c.** Si ripeta la procedura del punto a. per $M = 200$ iterazioni Monte Carlo. Chiamiamo $\hat{q}'_{\alpha,i}$ lo stimatore ottenuto all'i-ma iterazione. Si calcoli l'EQM sotto radice $(\sum_{i=1}^{M} (\hat{q}'_{\alpha,i} - q_\alpha)^2 / M)^{0.5}$, dove q_α è il vero quantile α di una distribuzione $\mathcal{N}(0,1)$. Si faccia poi la stessa cosa, ma usando la procedura del punto b.
  - **d.** Si commentino i risultati.

[5] In `R`, definito da `quantile(y, α)`, dove `y` è il vettore contenente i dati.