

Slide Set 7

Code, Outliers e Robustezza

Pietro Coretto
pcoretto@unisa.it

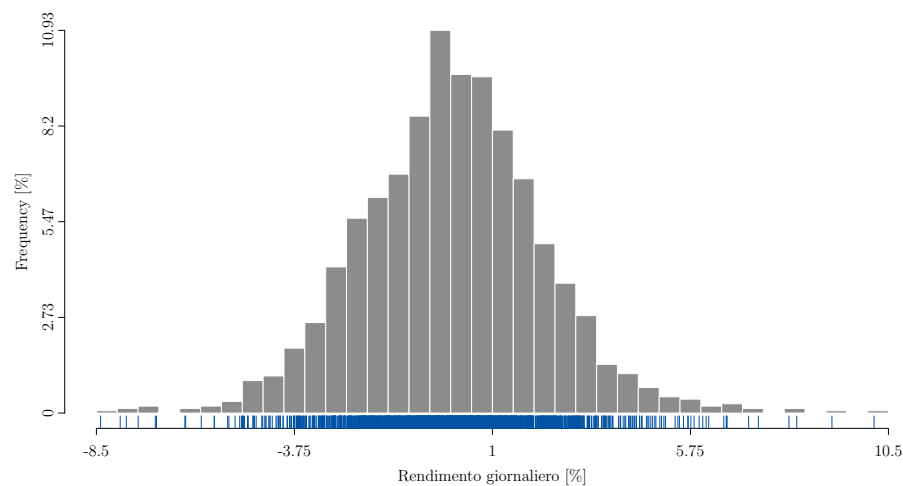
Corso di

Analisi e Visualizzazione dei Dati (Parte I)

Corso di Laurea in "Statistica per i Big Data" (L-41)
Università degli Studi di Salerno

Versione: 23 marzo 2022 (h15:17)

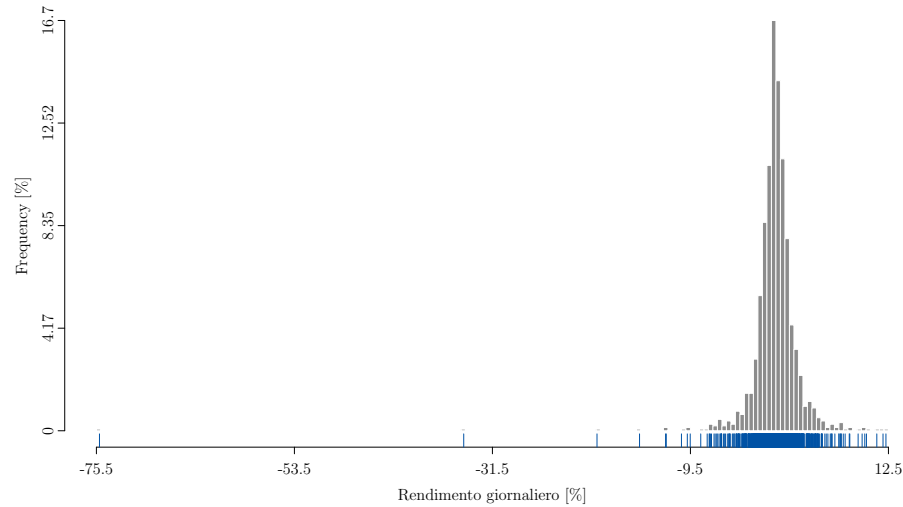
Dati: rendimenti (%) giornalieri, per "Nabors Industries Ltd" (mercato S&P500), periodo di contrattazione 2002-2007, $n = 1509$



Notes

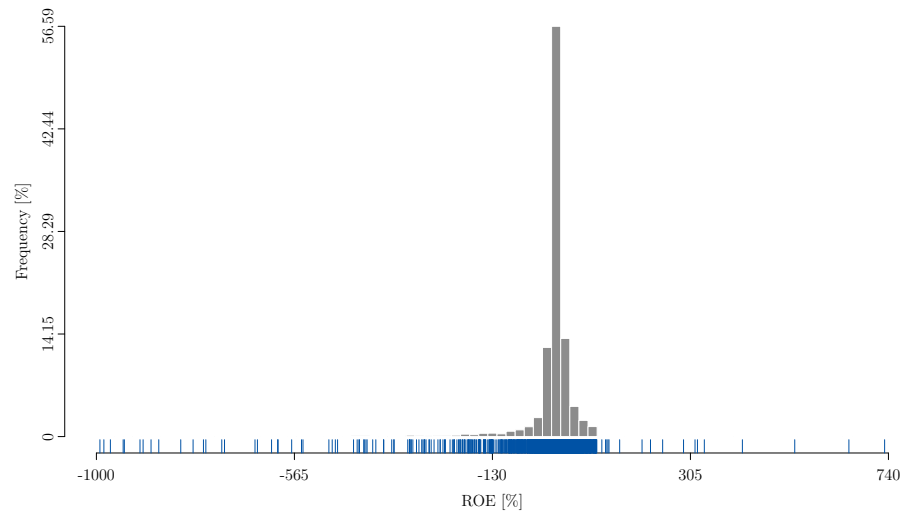
Notes

Dati: rendimenti (%) giornalieri, per “EDS” (mercato S&P500), periodo di contrattazione 2002-2007, $n = 1509$



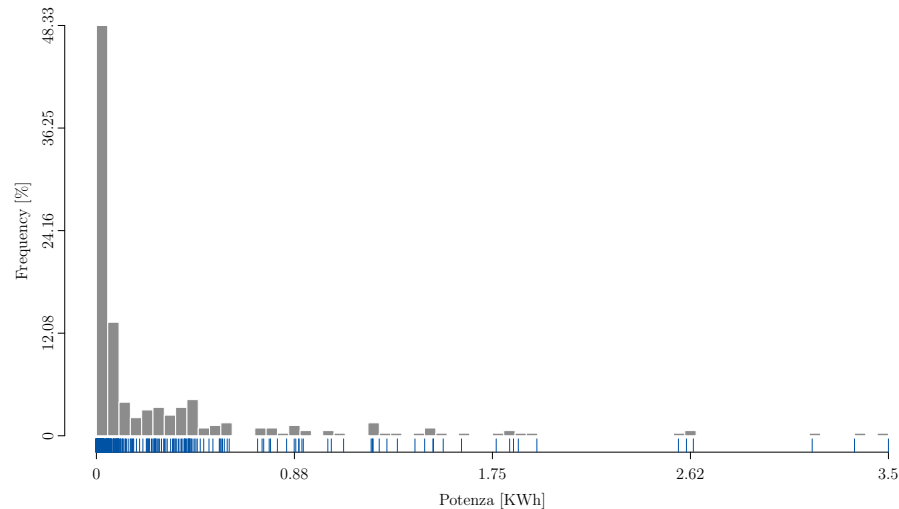
Notes

Dati: rendimenti (%) del capitale proprio (ROE), su $n = 5000$ imprese italiane



Notes

Dati: assorbimento della potenza elettrica su $n = 329$ contattori, utenze domestiche, fascia oraria 10:30-11:30



Notes

Code di una distribuzione

è la regione dove si collocano i dati che stanno *lontano* dalla massa centrale. Le code contengono i dati meno frequenti e più estremi

Domanda 1: perché ci interessano le code?

- osservazioni con livelli estremi (anche se sono poche) \Rightarrow problemi di visualizzazione
- sintesi numeriche (centralità, dispersione, etc) possono diventare poco rappresentative perché fortemente influenzate da una minoranza di punti nelle code

Domanda 2: perché certi dati hanno code così lunghe/estreme?

- questioni intrinseche al processo generatore dei dati (es: potenza, rendimenti)
- artifici introdotti dal meccanismo di misurazione e/o raccolta dati (es: ROE?)

Notes

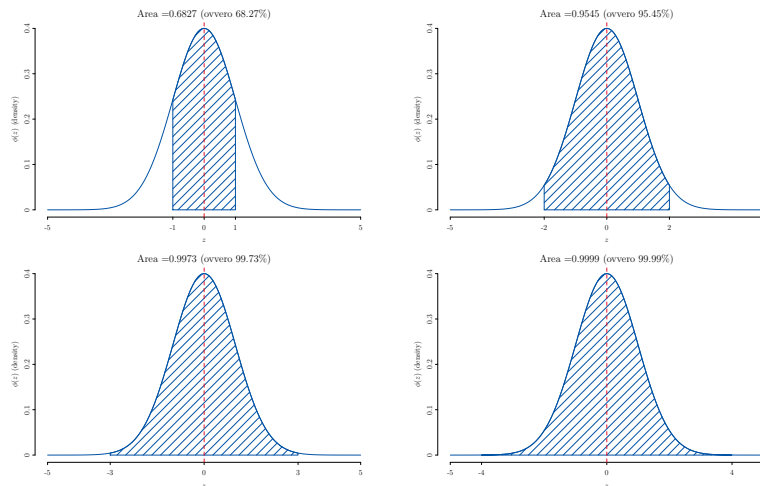
Curtosi

è una proprietà relativa al velocità con cui la densità/frequenza dei dati diminuisce mano mano che ci allontaniamo dal centro andando verso i valori più estremi

La curtosi è una nozione relativa. Valutazioni circa le code richiedono un riferimento, ovvero un termine di paragone

Assumiamo come riferimento il modello normale standard. Cosa accade se ci spostiamo dal centro verso i valori più estremi?

Notes



Notes

La **curtosi** del modello **normale** standard è caratterizzata dal **decadimento esponenziale** delle code rispetto alla deviazione quadratica dalla media

Osserva: per dati normali standard, ci aspettiamo $\bar{z} = 0$, quindi il quadrato

$$z_i^2 = (z_i - 0)^2 = (z_i - \bar{z})^2$$

coincide con la deviazione quadratica dalla media

Quando z_i diventa estremo la densità si annulla, infatti

$$\lim_{z_i \rightarrow \pm\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} = 0$$

Nelle code la densità ϕ si annulla ad una velocità paragonabile a

$$e^{-t} \rightarrow 0 \quad \text{per} \quad t \rightarrow \infty$$

dove t rappresenta z_i^2

Notes

Terminologia: quando abbiamo code dense, spesso diciamo

- **code spesse/pesanti:** code più ricche di dati del solito (si, ma più ricche di cosa?)
- **code lunghe:** presenza di dati a grande distanza dal bulk (si, ma a che distanza?)

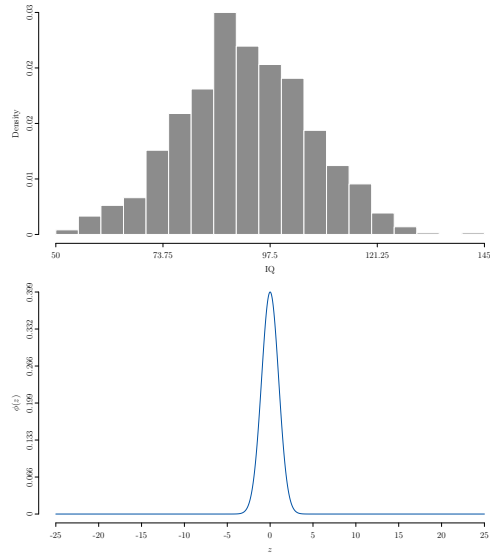
Tipologie di code vs Normale Standard

- **Platicurtiche**
code meno spesse/pesanti/lunghe rispetto alla normale standard
- **Mesocurtiche o Normocurtiche**
code spesse/pesanti/lunghe come quelle della normale standard
- **Leptocurtiche**
code più spesse/pesanti/lunghe rispetto alla normale standard

Valutazione della curtosi: data una variabile osservata X confronto il suo comportamento di coda rispetto al modello normale standard

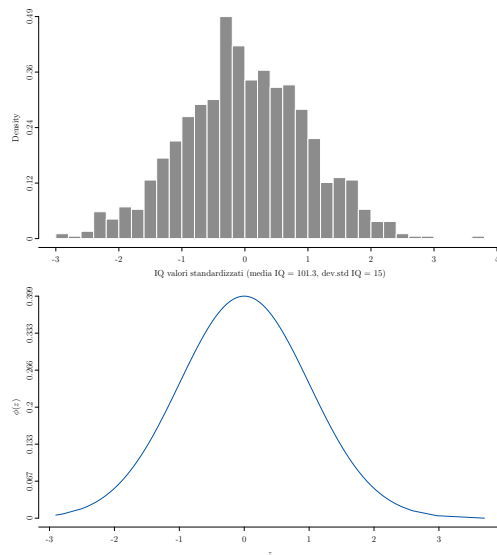
Notes

Paragonare le code di distribuzioni “*non allineate*” rispetto al loro centro/scaling, non è una buona idea. Dati: IQ



Notes

... standardizzando i valori di IQ



Notes

Le code di X sono le regioni dove si collocheranno i quantili più *esterni*, ovvero quelli calcolati con α lontano da $\alpha = 50\%$. Ad esempio $\alpha = 1\%$, 2% , 99% , 98% , \dots

Supponiamo che $\bar{x} = 0$ e $s_X^2 = 1$. X_α è un quantile esterno di X , mentre Z_α è il quantile implicato dal modello normale. Cosa mi aspetto?

- **Platicurtiche**

X_α è meno estremo di Z_α

- **Mesocurtiche o Normocurtiche**

X_α è approssimativamente uguale a Z_α

- **Leptocurtiche**

X_α è più estremo di Z_α

Normal Quantile-Quantile plot (Normal Q-Q plot)

Supponiamo di voler confrontare le code della X osservata con quelle del modello normale standard. Sia $\{x_1, x_2, \dots, x_n\}$ il campione osservato e sia Y la standardizzata di X , ovvero

$$y_i = \frac{x_i - \bar{x}}{s_X}, \quad \text{dove quindi } \bar{y} = 0 \text{ e } s_Y = 1$$

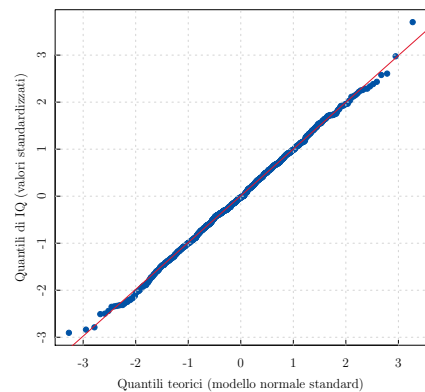
Costruzione del Normal Q-Q plot

- si fissa la griglia di valori per α , solitamente: $\alpha = \frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n}$
- si calcolano i quantili di Y per tutti i valori di α
- si calcolano i quantili nella normale standard (detti “*quantili teorici*”), per gli stessi valori di α
- si costruisce un semplice grafico dei quantili di Y vs i “*quantili teorici*”

Notes

Notes

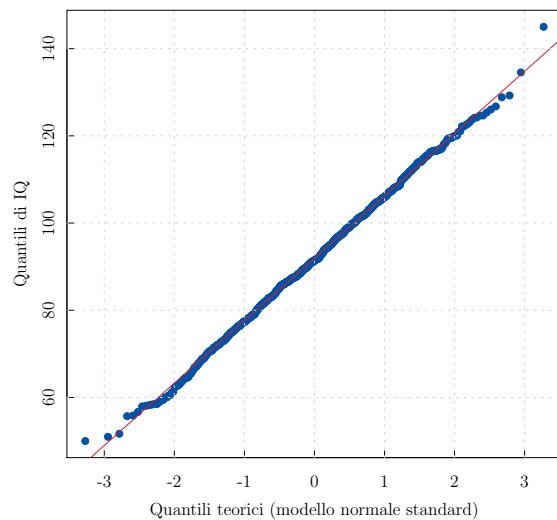
Dati IQ



Interpretazione: distribuzione approssimativamente normocurtica. Infatti i quantili di IQ si allineano piuttosto uniformemente a quelli che mi aspetterei sotto il modello normale

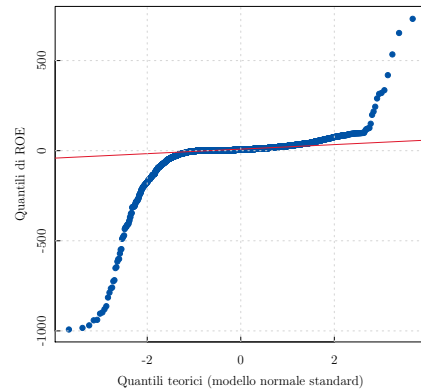
Notes

È possibile vedere la scala originale sull'asse verticale



Notes

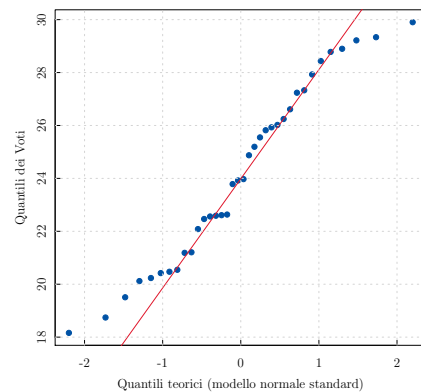
Dati ROE



Interpretazione: distribuzione chiaramente leptocurtica. Infatti: (i) sulla coda Sx i quantili sono più estremi della normale (quantili ROE < quantili normale); (i) sulla coda Dx i quantili sono più estremi della normale (quantili ROE > quantili normale)

Notes

Voti esame di Statistica del 25/01/2015:

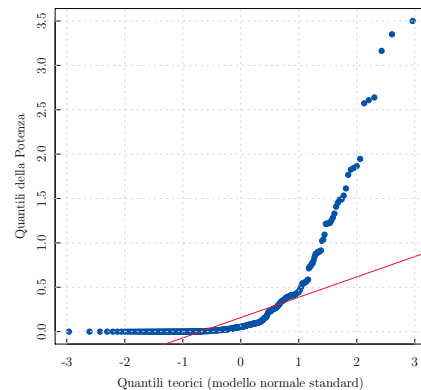


Interpretazione: distribuzione moderatamente platicurtica. Infatti: (i) sulla coda Sx i quantili sono meno estremi della normale (quantili Voti > quantili normale); (i) sulla coda Dx i quantili sono meno estremi della normale (quantili Voti < quantili normale)

Notes

⚠ Il Normal Q-Q plot non consente solo di valutare le code. . .

Dati: assorbimento della potenza elettrica



Interpretazione: la distribuzione della potenza devia fortemente dalla forma normale, si vede chiaramente una forte asimmetria positiva. Infatti: (ii) nella parte centrale il comportamento devia dalla normalità; (ii) la coda Dx molto più spessa della normale (iii) la coda Sx molto meno spessa della normale

Notes

Quantile-Quantile plot (Q-Q plot)

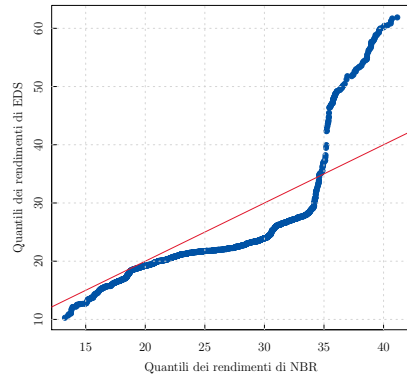
Supponiamo di osservare due variabili U e W . Supponiamo di voler confrontare le loro distribuzioni usando l'idea del Normal Q-Q-plot.

Costruzione del Q-Q plot

- si fissa la griglia di valori di α , solitamente: $\alpha = \frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n}$
- si calcolano i quantili della standardizzata di U per tutti i valori di α
- si calcolano i quantili della standardizzata di W per tutti i valori di α
- si costruisce un semplice grafico dei quantili di U vs i quantili di W

Notes

Dati: tassi di rendimento giornaliero (%), confronto della distribuzione dei rendimenti di NBR vs i rendimenti di EDS



Interpretazione: I rendimenti di EDS hanno code molto più spesse di NBR, con una differenza enorme sulla coda Dx. Quindi i rendimenti di EDS si caratterizzano per una sovrabbondanza di deviazioni sopra la media

Esempi/Applicazioni → [R script file](#)

Notes

Contaminazione dei dati e code pesanti

Contaminazione dei dati

i dati sono misurati e/o acquisiti con errori grossolani. Esempi:

- si misura l'altezza di un individuo in metri, invece di leggere/registrare 1.8 si registra 180
- si misura la pressione di un gas ad una certa temperatura, ma il sistema refrigerante del laboratorio è guasto

La condizione comune in questi esempi è che l'esistenza di dati “*anomali*” è determinata da fattori estranei al meccanismo generatore dei dati.

Outliers

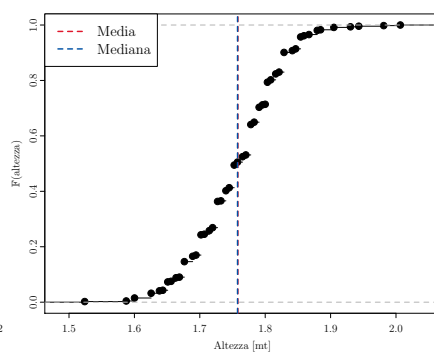
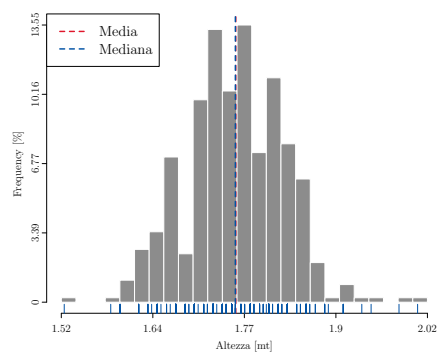
dati estranei al meccanismo che governa X , essi sono anomali / atipici / disomogenei rispetto al bulk

Notes

Dati: "Galton's height data", $n = 898$ soggetti

Altezza media = 1.7583mt

Altezza Mediana = 1.7577mt

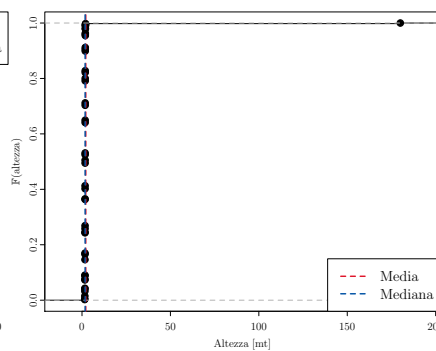
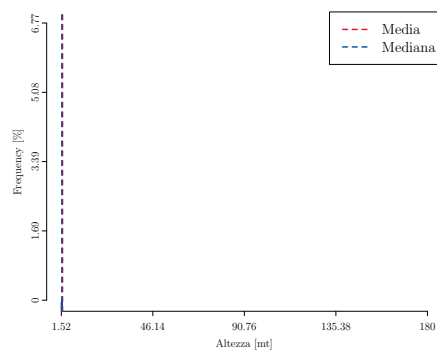


Notes

Supponiamo che in un caso 1.8mt sia stata registrata come 180

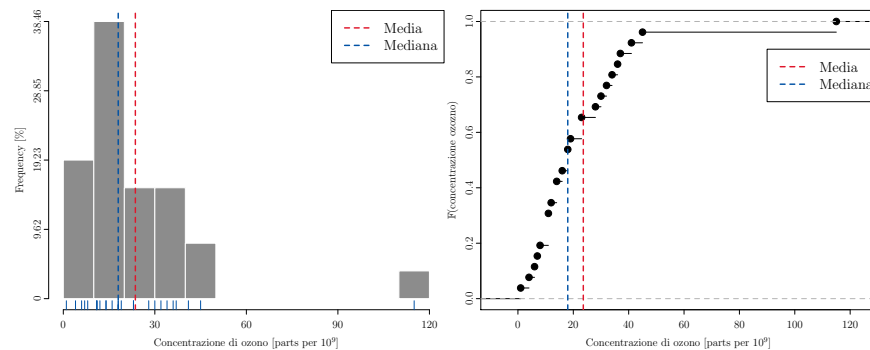
Altezza media = 2.1415mt

Altezza Mediana = 1.7577mt



Notes

Dati: "Air Quality Data", $n = 31$ misurazioni della concentrazione di varie sostanze nell'aria in Roosevelt Island (NYC, USA).



Notes

Gli outliers giacciono nelle code

gli outliers si presentano come valori grandi/piccoli rispetto al bulk, quindi rendono le code dei dati osservati pesanti

Un punto lontano dal *bulk* è sempre un outlier?

Risposta: assolutamente NO. Talvolta le code pesanti sono dovute alla natura del fenomeno (es: certi rendimenti finanziari). La difficoltà di distinguere le due cose è tanto maggiore quanto minore è n .

Esistono metodi per identificare gli outliers con certezza?

Risposta: assolutamente NO. Esistono metodi detti "*robusti*" che consentono di

- ridurre l'influenza delle code pesanti
- identificare punti che sono *potenziali* outliers

Notes

Robustezza

è la proprietà di *stabilità* di una procedura di analisi il cui risultato non cambia troppo quando una parte dei dati osservati è sottoposta ad una perturbazione

I metodi robusti sono utili per

- tenere sotto controllo l'effetto di outliers
- identificare i *potenziali* outliers
- in dati non contaminati, ma con code pesanti, tenere sotto controllo l'eccessiva influenza delle code quando n è modesto

Cosa qualifica la *robustezza* di un metodo? Esistono varie misure di robustezza

Notes

Definizione (Finite Sample Replacement Breakdown Point, *informale*)

Il breakdown point (BP) di una procedura di analisi è pari alla minima frazione di dati osservati che arbitrariamente perturbati rendono i risultati dell'analisi privi di qualsiasi senso

Osserva

- Il BP è una misura di *stabilità* rispetto a perturbazioni dei dati osservati
- BP=basso \implies procedure molto *sensibili* alle perturbazioni dei dati
- BP=alto \implies procedure molto *resistenti* alle perturbazioni dei dati
- Se gli outliers dovessero superare il 50% di n , questi costituirebbero la maggioranza. Quindi BP=50% è considerato il massimo desiderabile.

Notes

Proprietà (BP di media e mediana)

Sia $\{x_1, x_2, \dots, x_n\}$ il campione osservato per una variabile X :

- 1 il BP della media è $\frac{1}{n}$
- 2 il BP della mediana è $\frac{1}{n} \lfloor \frac{n+1}{2} \rfloor$

Dimostrazione. Consideriamo prima il caso della media

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_{n-1}}{n} + \frac{x_n}{n}$$

cambiamo arbitrariamente $x_n \rightarrow \pm\infty$, di conseguenza $\bar{x} \rightarrow \pm\infty$. Quindi perturbando arbitrariamente 1 sola osservazione su n , \bar{x} lascia il range dei dati arbitrariamente. Quindi il BP di \bar{x} è $1/n$.

Notes

Passiamo alla mediana. Consideriamo le $h = \lfloor \frac{n}{2} \rfloor$ più piccole della mediana

$$\underbrace{x_{(1)}, x_{(2)}, \dots, x_{(h)}}_{\lfloor \frac{n}{2} \rfloor \text{ osservazioni}} < X_{0.5} \leq \underbrace{x_{(h+1)}, x_{(h+2)}, \dots, x_{(n)}}_{n - \lfloor \frac{n}{2} \rfloor \text{ osservazioni}}$$

Perturbiamo arbitrariamente $x_{(1)}, x_{(2)}, \dots, x_{(h)} \rightarrow -\infty$, la mediana non può cambiare arbitrariamente perché $X_{0.5}$ è ancora nella maggioranza delle osservazioni (non perturbate).

Tuttavia, se la perturbazione riguardasse anche $x_{(h+1)} \rightarrow -\infty$, allora $X_{0.5} \rightarrow -\infty$

In conclusione devo perturbare almeno $\lfloor \frac{n+1}{2} \rfloor$ dati per far spostare la mediana arbitrariamente, quindi la frazione di breakdown è $\frac{1}{n} \lfloor \frac{n+1}{2} \rfloor$

Il ragionamento sulla coda destra dei dati condurrebbe alla stessa conclusione



Notes

Osserva

- media: $BP = \frac{1}{n} \rightarrow 0$ quando $n \rightarrow \infty$
- mediana: $BP = \frac{1}{n} \left\lfloor \frac{n+1}{2} \right\rfloor \rightarrow \frac{1}{2}$ (50%) quando $n \rightarrow \infty$
- per la mediana già con $n = 20$ otteniamo un $BP \approx 0.5$

Misure di posizione/centralità robuste

- Mediana
- Trimean
- Medie troncate
- etc

Misure di dispersione robuste

- IQR
- MAD
- etc

Notes

Trimean (BP=25% per n grande)

è un somma pesata di mediana e quartili

$$TM = \frac{1}{2} \text{Med} + \frac{1}{4} Q_1 + \frac{1}{4} Q_3 \quad (7.1)$$

A differenza della mediana rappresenta meglio la parte centrale di distribuzioni asimmetriche

Medie troncate di livello α (BP= $\alpha\%$ per n grande)

media sulle $(1 - \alpha)n$ osservazioni più centrali ottenute troncando le αn osservazioni più estreme:

- $M = \lceil n \frac{\alpha}{2} \rceil$: numero di punti da troncare su ciascuna coda
- ordinati i dati, si calcola la media delle $n - 2M$ osservazioni centrali

$$\bar{x}_\alpha = \frac{1}{n - 2M} \sum_{j=1+M}^{n-M} x_{(j)} \quad (7.2)$$

Notes

Median absolute deviation, noto come MAD (BP=50% per n grande)


è una misura di dispersione definita come la mediana delle deviazioni assolute dalla mediana:

- $\text{Med}(X) = \text{Med}(\{x_1, x_2, \dots, x_n\})$
- per ogni x_i calcoliamo la deviazione assolute dalla mediana δ_i ovvero

$$\delta_1 = |x_1 - \text{Med}(X)|, \dots, \delta_n = |x_n - \text{Med}(X)|$$

- calcoliamo il MAD come la mediana delle deviazioni

$$\text{Mad}(X) = \text{Med} \{ \delta_1, \delta_2, \dots, \delta_n \}$$

 per questioni che non possiamo introdurre in questo corso, il MAD viene solitamente moltiplicato per una costante. Solitamente

$$\text{Mad}(X) = 1.4826 \times \text{Med} \{ \delta_1, \delta_2, \dots, \delta_n \}$$

Notes

Considerazioni finali sulla robustezza

- la robustezza è una questione fondamentale soprattutto nell'era dei Big Data
- la letteratura sulle tecniche *robuste* ad alto BP è enorme
- un alto BP si ottiene solo riducendo il peso attribuito alle code (pensiamo alla media troncata). Questo ha un costo che sarà chiaro nei corsi successivi
- sebbene le procedure ad alto BP nascono per risolvere problemi di contaminazione/outliers, esse sono utilizzate anche per *stabilizzare* l'analisi in caso di code pesanti e poche osservazioni (n modesto)

Esempi/Applicazioni → [R script file](#)

Notes

Si stabilisce una regola di classificazione in base alla quale un punto viene “sospettato di essere un outlier”

La regola si deve basare su strumenti di provata robustezza, altrimenti la regola stessa sarebbe affetta da contaminazione

Non ci sono essere certezze su cosa sia “*veramente estraneo ai dati*”, soprattutto quando n è modesto. L'identificazione deve essere piuttosto interpretata come l’“*attribuzione di un flag ad un dato sospetto*” (J. W. Tukey, 1977)

Tukey fences

regola di identificazione semplicissima introdotta da Tukey, si basa semplicemente sui quartili e Q_1 , Q_3 ed il corrispondente IQR.

Step 1:

sui dati $\{x_1, x_2, \dots, x_n\}$, calcoliamo Q_1 e Q_3 , e $\text{IQR} = Q_3 - Q_1$

Step 2:

fissato un numero $k > 1$, calcoliamo l'intervallo $[H_{\text{inf}}, H_{\text{sup}}]$ dove

$$H_{\text{inf}} = Q_1 - k \text{IQR}$$

$$H_{\text{sup}} = Q_3 + k \text{IQR}$$

questi sono detti “fences” (recinti) nel linguaggio di Tukey

Step 3:

i **dati sospetti** sono quelli esterni al “*recinto*” $[H_{\text{inf}}, H_{\text{sup}}]$

Classificazione

- **Punti regolari**: le osservazioni che stanno dentro

$$[Q_1 - 1.5 \text{ IQR}, Q_3 + 1.5 \text{ IQR}] \quad (k = 1.5)$$

- **(sospetto) outlier**: l'osservazione che sta esternamente a

$$[Q_1 - 1.5 \text{ IQR}, Q_3 + 1.5 \text{ IQR}] \quad (k = 1.5)$$

- **(sospetto) gross outlier**: l'osservazione che sta esternamente a

$$[Q_1 - 3 \text{ IQR}, Q_3 + 3 \text{ IQR}] \quad (k = 3)$$

Domanda: perché $k = 1.5$ e $k = 3$? Si dimostra che se i dati avessero una distribuzione simile a quella normale allora

- circa 97.5% delle osservazioni centrali starebbero dentro

$$[Q_1 - 1.5 \text{ IQR}, Q_3 + 1.5 \text{ IQR}]$$

- circa 99.99% delle osservazioni centrali starebbero dentro

$$[Q_1 - 3 \text{ IQR}, Q_3 + 3 \text{ IQR}]$$

Notes

Esempio: si osserva

$$\{-10, 10, 11, 5, 8, 9, 16\}$$

da cui $Q_1 = 6.5$, $Q_2 = \text{Med} = 9$, $Q_3 = 10.5$, e $\text{IQR} = Q_3 - Q_1 = 4$

Calcoliamo i “fences” per identificare outliers:

$$[Q_1 - 1.5 \text{ IQR}, Q_3 + 1.5 \text{ IQR}] = [0.15, 16.5]$$

sono sospetti outliers i punti esterni, quindi $X = -10$

Calcoliamo i “fences” per identificare i gross outliers:

$$[Q_1 - 3 \text{ IQR}, Q_3 + 3 \text{ IQR}] = [-5.5, 22.5]$$

sono sospetti gross outliers i punti esterni, quindi ancora una volta $X = -10$

Esempi/Applicazioni → **R script file**

Notes