Modelli descrittivi, statistica e simulazione

Master per Smart Logistics specialist

Roberto Cordone
(roberto.cordone@unimi.it)

Statistica inferenziale — Cernusco S.N., giovedì 25 febbraio 2016 (9.00/13.00)

Inferenza statistica

Sinora si è supposto sempre di conoscere l'intera popolazione

- questa ipotesi di solito è falsa
- molto spesso si conosce solo un campione, per motivi di
  - costo (differenza tra sondaggio e censimento)
  - possibilità fisica (la popolazione contiene individui inaccessibili, per es. esiti futuri)

L'inferenza statistica consiste nel indurre le proprietà di un'intera popolazione in base ai dati conosciuti su un suo campione

Richiede tre fasi

1. estrazione di un campione dalla popolazione
2. descrizione del campione, cioè calcolo dei suoi indici statistici
3. formulazione delle inferenze sull'intera popolazione

Estrazione del campione

L'estrazione del campione è un'operazione importante

- il campione deve essere
  - abbastanza grande
  - scelto casualmente, dando ugual probabilità a ogni individuo
- un campione "viziato" o polarizzato porta a inferenze scorrette
  Esempio: i sondaggi telefonici descrivono solo la sottopopolazione che è a casa nelle ore scelte per il sondaggio ed è disponibile a rispondere

La descrizione del campione avviene come discusso nelle lezioni sulla statistica descrittiva

Formulazione di inferenze

L'idea fondamentale è

- introdurre delle variabili aleatorie che descrivono caratteristiche interessanti
- studiare la distribuzione di tali variabili sul campione
- chiedersi se tale distribuzione somiglia a quella che assumono sull'intera popolazione

Più specificamente, questo interrogarsi può assumere le seguenti forme

- stima puntuale: stimare un valore per un parametro p della distribuzione della popolazione (media, varianza, ecc. . .)
- intervallo di confidenza: stimare un intervallo di valori in cui p rientra con probabilità superiore a una data soglia
- test di ipotesi: stimare la probabilità con cui p rispetta una data ipotesi sul proprio valore (per es., essere inferiore a una data soglia)

Esempio

La distribuzione degli esiti dei lanci per un dado non è esattamente nota (è uniforme se il dado è non truccato, ma lo è davvero?)

1. Estrazione del campione: lanciamo il dado un certo numero di volte
2. Studio della distribuzione: calcoliamo la media campionaria
3. Inferenza: possiamo fare due cose complementari
   - stimare un intervallo nel quale la media della popolazione rientra con alta probabilità e vedere se la media teorica (3.5) vi rientra
   - valutare la probabilità che la media sia quella teorica in base ai dati

Come funziona tutto questo?

Stimatori

Gli indici di un campione estratto casualmente

- dopo l'estrazione, sono valori numerici deterministici
- prima dell'estrazione, sono variabili aleatorie,
  dato che dipendono dal campione effettivamente estratto

Stimatore è una variabile aleatoria che descrive i possibili valori di un indice statistico associato a un campione casuale da estrarre

Come esempio di indice, consideriamo la media

- ogni campione possibile è un esito di un esperimento casuale
- la media del campione estratto è il valore associato a quell'esito
- la media campionaria è una variabile aleatoria con una sua distribuzione statistica
- la funzione di distribuzione $F(t)$ di tale variabile fornisce la probabilità che la media sia $\leq t$
- da questa probabilità possiamo ricavare informazioni utili

Esempio

Se lanciando una moneta non truccata "testa" corrisponde a 0 e "croce" a 1,

- il risultato è una variabile aleatoria con distribuzione uniforme in $\{0, 1\}$
- la sua media è $\mu = \dfrac{0+1}{2} = 0.5$

Se estraiamo un campione di n individui, cioè facciamo n lanci,
la media campionaria $\bar{x} = \sum_{j=1}^{n} \dfrac{x_j}{n} = \dfrac{n_1}{n}$ non è sempre 0.5

- può essere 0 (se esce testa ad ogni lancio), ma è improbabile
- può essere 1 (se esce croce ad ogni lancio), ma è improbabile
- può assumere valori intermedi, con un picco di probabilità in 0.5

Com'è la distribuzione della media campionaria?

A questa domanda sappiamo rispondere

Esempio

Com'è la distribuzione della media campionaria?

La media campionaria è semplicemente $\bar{x} = \frac{n_1}{n}$, il numero di croci diviso n

Il numero di croci segue la distribuzione binomiale, $\text{Bin}(n, 0.5)$

Questo significa che possiamo rispondere a domande del tipo

- che probabilità c'è che la media sia $\bar{x} \leq 0.2$?
- che probabilità c'è che la media sia $\bar{x} \geq 0.5$?
- che probabilità c'è che la media sia $\bar{x} \in [0.43; 0.55]$?

Se la media osservata è $\bar{x} = 0.1$ e la probabilità che sia ≤ 0.1 è minima,
è lecito il ragionevole sospetto che la moneta sia truccata per favorire la testa

Stima di indici statistici

Il valore di un indice statistico valutato su un campione (stimatore)

- per alcuni campioni, è vicino al valore dello stesso indice sull'intera popolazione
- per altri campioni, il valore è lontano

Quando il valore campionario è lontano dal valore ipotizzato

1. o il campione è un po' particolare
2. o l'ipotesi fatta sulla distribuzione della popolazione è errata

Se la prima spiegazione ha probabilità bassa, si può optare per la seconda

Ci concentriamo sulla media campionaria, che è uno stimatore della media

$$\bar{x} = \sum_{i \in C} \frac{x_i}{|C|}$$

Media campionaria

Come tutte le variabili aleatorie, la media campionaria ha indici statistici (media, varianza, ecc. . .)

Si dimostra che la media campionaria di qualsiasi variabile aleatoria ha

- media uguale alla media della variabile di partenza
- deviazione standard uguale alla deviazione standard della variabile di partenza divisa per $\sqrt{n}$
- distribuzione che tende alla distribuzione normale per $n \to \infty$ (teorema del limite centrale)

Poiché media e deviazione standard identificano la distribuzione normale, la distribuzione della media campionaria su campioni abbastanza grandi è approssimativamente nota per ogni variabile aleatoria

Questo aiuta a giudicare se la media campionaria osservata è verosimile come stima della media per l'intera popolazione, oppure non lo è

Stima puntuale

Esistono tre approcci, basati sugli stessi concetti, ma con risultati diversi

1. stima puntuale
2. intervallo di confidenza
3. test di ipotesi

Nel caso della media, la stima puntuale consiste semplicemente nel

- valutare la media campionaria $\bar{x} = \sum_{i \in C} \frac{x_i}{|C|}$
- usare $\sigma/\sqrt{n}$ come misura di accuratezza della stima,
  cioè l'errore cresce o cala secondo che $\sigma/\sqrt{n}$ cresce o cala
  (σ è la deviazione standard della variabile originale)

Regione di rifiuto

Come si è detto, il valore di uno stimatore

- per alcuni campioni, è vicino al valore dell'indice stimato
- per altri campioni, il valore è lontano

Regione di rifiuto è l'insieme dei campioni con indici di valore improbabile

Per convenzione, si considerano non plausibili i fenomeni la cui probabilità complessiva di manifestarsi è $\leq 5\%$ (se si è più conservativi, 1%)

Si vuole costruire un insieme (di solito, un intervallo) tale che i campioni la cui media cade fuori dall'insieme formino una regione di rifiuto

Esistono proprietà generali che consentono di stimare un'approssimazione della regione di rifiuto senza studiare ogni volta il problema specifico

Intervalli di confidenza

La media campionaria si comporta come una variabile normale di media μ e deviazione standard $\sigma/\sqrt{n}$ (teorema del limite centrale)

La variabile aleatoria ottenuta da $\bar{x}$ sottraendo μ e dividendo per $\sigma/\sqrt{n}$ ha distribuzione normale standard (media 0 e deviazione standard 1)

$$P\left[\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq t\right] = F_{N(0,1)}(t)$$

dove i valori di $F_{N(0,1)}(t)$ sono noti da tabelle o funzioni Excel

A questo punto

$$P\left[t_1 \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq t_2\right] = F_{N(0,1)}(t_2) - F_{N(0,1)}(t_1)$$

e quindi

$$P\left[\bar{x} - t_2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} - t_1\frac{\sigma}{\sqrt{n}}\right] = P\left[t_1 \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq t_2\right] = F_{N(0,1)}(t_2) - F_{N(0,1)}(t_1)$$

Che ne ricaviamo?

Intervalli di confidenza

Ne ricaviamo la probabilità che la media μ cada molto sopra, molto sotto o molto lontano dalla media campionaria $\bar{x}$

$$\begin{aligned} P\left[\mu \leq \bar{x} - t\frac{\sigma}{\sqrt{n}}\right] &= F_{N(0,1)}(t) \\ P\left[\mu \geq \bar{x} + t\frac{\sigma}{\sqrt{n}}\right] &= F_{N(0,1)}(t) \\ P\left[\bar{x} - t\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + t\frac{\sigma}{\sqrt{n}}\right] &= 2F_{N(0,1)}(t) \end{aligned}$$

Sia α l'errore massimo accettabile nella stima della media

- si impone $P\left[\mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}}\right] \geq 1 - \alpha$, fissando $t \geq F^{-1}_{N(0,1)}(\alpha)$
- si impone $P\left[\mu \geq \bar{x} - \frac{\sigma}{\sqrt{n}}\right] \geq 1 - \alpha$, fissando $t \geq F^{-1}_{N(0,1)}(\alpha)$
- si impone $P\left[|\bar{x} - \mu| \leq t\frac{\sigma}{\sqrt{n}}\right] \geq 1 - \alpha$, fissando $t \geq F^{-1}_{N(0,1)}(\alpha/2)$

Il valore di $F^{-1}_{N(0,1)}(\alpha)$, spesso detto Z_α, si ricava da tabelle o Excel

Intervalli di confidenza con varianza incognita

Il metodo richiede il valore della deviazione standard σ della popolazione

Se σ è incognita, usare il valore campionario con $\bar{x}$ al posto di μ

$$S_n^2 = \sum_{i \in C} \frac{(x_i - \bar{x})^2}{n}$$

porta ad ottenere un intervallo di confidenza approssimato

La strategia teoricamente corretta però è

- usare la varianza campionaria corretta

$$S_{n-1}^2 = \sum_{i \in C} \frac{(x_i - \bar{x})^2}{n-1}$$

  che tiene conto della sostituzione della media μ con la media campionaria $\bar{x}$, dividendo per $n-1$ anziché n (la varianza cresce)
- usare la distribuzione t di Student al posto della normale standard

Per campioni molto grandi, i due intervalli di confidenza differiscono poco

Intervalli di confidenza

Il significato pratico di intervallo di confidenza al 5% è:

- per almeno il 95% dei campioni l'intervallo $\left[\bar{x} - \frac{\sigma}{\sqrt{n}}, \bar{x} + \frac{\sigma}{\sqrt{n}}\right]$ contiene la media μ
- per al massimo il 5% dei campioni l'intervallo non contiene la media