

Calcolo delle Probabilità e Statistica

Teorema del Limite Centrale e Stima Parametrica

Ionel Eduard Stan¹

¹Dip. di Matematica e Informatica, Università di Ferrara ioneleduard.stan@unife.it

Discutiamo alcuni problemi fondamentali legati all'andamento asintotico di sequenze di variabili aleatorie. Si consideri una sequenza X_1, X_2, \dots di variabili aleatorie *indipendenti e identicamente distribuite (i.i.d)* con media μ e varianza σ^2 . Sia:

$$S_n = \sum_{i=1}^n X_i,$$

la somma delle prime n di esse. I teoremi limite sono principalmente interessati alle proprietà di S_n e le sue variabili aleatorie relazionate per n crescente.

Dall'indipendenza, abbiamo che:

$$\begin{aligned} \text{var}(S_n) &= \text{var}(\sum_{i=1}^n X_i) \\ &= \text{var}(X_1 + X_2 + \dots + X_n) \\ &= \text{var}(X_1) + \text{var}(X_2) + \dots + \text{var}(X_n) \\ &= \sum_{i=1}^n \text{var}(X_i) \\ &= n \cdot \sigma^2. \end{aligned}$$

Dunque, la distribuzione di S_n si diffonde all'aumentare di n , e non può avere un limite significativo. La situazione è differente se consideriamo la *media campionaria*:

$$M_n = \frac{\sum_{i=1}^n X_i}{n}.$$

Calcolando velocemente abbiamo che:

$$\begin{aligned}
 E[M_n] &= E\left[\frac{\sum_{i=1}^n X_i}{n}\right] \\
 &= \frac{1}{n} \cdot E[\sum_{i=1}^n X_i] \\
 &= \frac{1}{n} \cdot E[X_1 + X_2 + \cdots + X_n] \\
 &= \frac{1}{n} \cdot (E[X_1] + E[X_2] + \cdots + E[X_n]) \\
 &= \frac{1}{n} \cdot \sum_{i=1}^n E[X_i] \\
 &= \frac{n \cdot \mu}{n} \\
 &= \mu,
 \end{aligned}$$

e

$$\begin{aligned}
 \text{var}(M_n) &= \text{var}\left(\frac{\sum_{i=1}^n X_i}{n}\right) \\
 &= \text{var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\
 &= \frac{1}{n^2} (\text{var}(X_1) + \text{var}(X_2) + \cdots + \text{var}(X_n)) \quad (\text{indipendenza}) \\
 &= \frac{1}{n^2} \cdot \sum_{i=1}^n \text{var}(X_i) \\
 &= \frac{n\sigma^2}{n^2} \\
 &= \frac{\sigma^2}{n}.
 \end{aligned}$$

In particolare, la varianza di M_n va a 0 per n crescente, e la distribuzione di M_n è centrata su μ .

Considereremo anche una quantità intermedia tra S_n e M_n :

$$Z_n = \frac{S_n - n \cdot \mu}{\sigma \cdot \sqrt{n}}.$$

Si può dimostrare che:

$$E[Z_n] = 0, \quad \text{var}(Z_n) = 1.$$

Siccome la media e la varianza di Z_n rimangono invariate per n crescente, la sua distribuzione né si diffonde né si riduce ad un punto. Il *teorema del limite centrale* si focalizza sulla forma assintotica della distribuzione di Z_n e asserisce che diventa la distribuzione normale standard.

Definizione 1 (Campione aleatorio). Se X_1, X_2, \dots, X_n sono variabili aleatorie indipendenti tutte con la stessa distribuzione \mathcal{F} , allora diciamo che loro sono un campione aleatorio (random sample, in inglese), o semplicemente campione, dalla distribuzione \mathcal{F} . \blacklozenge

I problemi per i cui la forma della distribuzione è nota fino ad un insieme di parametri sconosciuti si dicono problemi di inferenza *parametrica*, mentre quelli in cui non sappiamo nulla sulla distribuzione si chiamano problemi di inferenza *non parametrica*.

Ci interessiamo delle distribuzioni di probabilità di certe statistiche che emergono dal campione, dove una *statistica* è una variabile aleatoria il cui valore è determinato dal campione.

In generale, consideriamo una *popolazione* di elementi, ognuno dei quali hanno un valore numerico (e.g., età, altezza, ecc.).

Teorema 1 (Legge "debole" dei grandi numeri). *Siano X_1, X_2, \dots variabili aleatorie i.i.d. con media μ . Per ciascuno $\epsilon > 0$, abbiamo che:*

$$P(|M_n - \mu| \geq \epsilon) = P\left(\left|\frac{X_1 + X_2 + \dots + X_n}{n} - \mu\right| \geq \epsilon\right) \rightarrow 0,$$

per $n \rightarrow \infty$. □

Intuitivamente, il teorema precedente asserisce che la media campionaria M_n di un numero grande di variabili aleatorie i.i.d. è molto vicina alla vera media, con alta probabilità.

Teorema 2 (Teorema del limite centrale). *Sia $S_n = X_1, X_2, \dots, X_n$ una sequenza di n variabili aleatorie i.i.d. ciascuna con media μ e varianza σ^2 . Allora, per n tendente ad infinito, la distribuzione di S_n è approssimativamente una normale con media $n \cdot \mu$ e varianza $n \cdot \sigma^2$.* □

Dal Teorema 1, la distribuzione della media campionaria $M_n = (X_1 + X_2 + \dots + X_n)/n$ si centra sempre di più nella vera media μ . In particolare, la varianza tende a 0. Dall'altra parte, la varianza della somma S_n non converge. Un risultato intermedio è di considerare la deviazione $S_n - n \cdot \mu$ di S_n dalla sua media $n \cdot \mu$, e scalare tale valore con un fattore proporzionale a $1/\sqrt{n}$. La cosa interessante di questa scalatura è che tiene la varianza ad un livello costante. Dal Teorema 2 sappiamo che la distribuzione di questa variabile aleatoria scalata si avvicina ad una normale. Sia, dunque, X_1, X_2, \dots una sequenza di variabili aleatorie i.i.d. con media μ e varianza σ^2 , e definiamo:

$$Z_n = \frac{X_1 + X_2 + \dots + X_n - n \cdot \mu}{\sigma \cdot \sqrt{n}}.$$

Allora, la CDF di Z_n converge ad una CDF normale standard, nel senso che:

$$\lim_{n \rightarrow \infty} P(Z_n \leq z) = \Phi(z),$$

per ciascun z .

Il Teorema del Limite Centrale ci permette di calcolare le probabilità di Z_n come se Z_n fosse una normale. Siccome le trasformazioni lineari preservano la normalità, questo è equivalente a trattare S_n come una variabile normale con media $n \cdot \mu$ e varianza $n \cdot \sigma^2$.

Abbiamo una procedura per l'approssimazione ad una normale basandoci sul Teorema del Limite Centrale. Sia $S_n = X_1 + X_2 + \dots + X_n$, dove le X_i sono variabili aleatorie i.i.d. con media μ e varianza σ^2 . Se n è sufficientemente grande, la probabilità $P(Z_n \leq c)$ può essere approssimata trattando S_n come una normale:

1. Calcolare la media $n \cdot \mu$ e la varianza $n \cdot \sigma^2$ di S_n .
2. Calcolare il valore normalizzato $z = (c - n \cdot \mu) / (\sigma \cdot \sqrt{n})$.
3. Usare l'approssimazione

$$P(S_n \leq c) \approx \Phi(z),$$

dove $\Phi(z)$ è disponibile nella tabella della CDF delle normali standard.

La legge "forte" dei grandi numeri è simile a quella "debole" per quanto riguarda la convergenza della media campionaria alla vera media. È differente, tuttavia, perché fa riferimento ad un altro tipo di convergenza.

Teorema 3 (Legge "forte" dei grandi numeri). *Sia X_1, X_2, \dots una sequenza di variabili aleatorie i.i.d con media μ . Allora, la sequenza della media campionaria $M_n = (X_1 + \dots + X_n) / n$ converge a μ , con probabilità 1, nel senso che:*

$$P\left(\lim_{n \rightarrow \infty} \frac{X_1 + \dots + X_n}{n} = \mu\right) = 1.$$

□

Per interpretare il teorema precedente, dobbiamo tornare alla descrizione dei modelli probabilistici in termini di spazi campionari. L'esperimento considerato è infinitamente lungo e genera una sequenza di valori, uno

per ciascuna variabile aleatoria della sequenza X_1, X_2, \dots . Dunque, possiamo immaginare lo spazio campionario come un'insieme di sequenze infinite (x_1, x_2, \dots) di numeri reali: ogni sequenza di questo tipo è il risultato dell'esperimento. Consideriamo adesso l'insieme A delle sequenze (x_1, x_2, \dots) la cui media nel lungo termine è μ , cioè:

$$(x_1, x_2, \dots) \in A \iff \lim_{n \rightarrow \infty} \frac{x_1 + \dots + x_n}{n} = \mu.$$

La legge "forte" dei grandi numeri dice che tutta la probabilità si concentra su questo sottoinsieme particolare dello spazio campionario. Equivalentemente, la collezione dei risultati che non appartengono ad A hanno probabilità 0.

La differenza tra la legge "debole" e quella "forte" dei grandi numeri è sottile e merita un'attenta esaminazione. La legge "debole" dice che la probabilità $P(|M_n - \mu| \geq \epsilon)$ va a zero per $n \rightarrow \infty$. Tuttavia, per qualsiasi n finito, questa probabilità può essere positiva ed è concepibile che una volta ogni tanto, anche se poco frequente, M_n devia significativamente da μ . La legge "debole" non fornisce nessuna informazione conclusiva sul numero di queste deviazioni, ma la legge "forte" sì. Secondo la legge "forte", e con probabilità 1, M_n converge a μ . Questo implica che per ciascun $\epsilon > 0$, la probabilità che la differenza $|M_n - \mu|$ supera ϵ un numero infinito di volte è uguale a zero.

Esempio 1. Prima di giocare alla roulette al casinò, potremmo essere interessati a pregiudizi (*biases*, in inglese) da sfruttare. Dunque, osserviamo 100 partite i cui risultati è un numero tra 1 e 36, e contiamo il numero di partite il cui risultato è dispari. Se il numero è maggiore di 55, allora decidiamo che la roulette non è equa. Assumendo che la roulette sia equa, trovare un'approssimazione alla probabilità che faremo la scelta sbagliata. \diamond

Esempio 2. Siano $X_1, Y_1, X_2, Y_2, \dots$ variabili aleatorie indipendenti, uniformemente distribuite nell'intervallo unitario $[0, 1]$, e sia:

$$W = \frac{(X_1 + \dots + X_{16}) - (Y_1 + \dots + Y_{16})}{16}.$$

Trovare un'approssimazione numerica alla quantità:

$$P(|W - E[W]| < 0.001).$$

◇

Adesso ci preoccupiamo della stima parametrica, usando l'approccio classico dove il parametro θ non è random, ma è visto come una costante sconosciuta.

Definizione 2 (Stimatore). *Date le osservazioni $X = (X_1, X_2, \dots, X_n)$, uno stimatore (estimator, in inglese) è una variabile aleatoria della forma $\hat{\Theta} = g(X)$, per qualche funzione g .* ♦

Si noti che, siccome la distribuzione di X dipende da θ , lo stesso vale anche per $\hat{\Theta}$. Usiamo il termine *stima* (estimate, in inglese) per fare riferimento ad un valore realizzato di $\hat{\Theta}$. In particolare, quando siamo interessati al ruolo delle n osservazioni, usiamo la notazione $\hat{\theta}_n$ per lo stimatore; si può interpretare come una sequenza di stimatori (uno per ciascun valore di n). La media e la varianza di $\hat{\Theta}_n$ si denotano con $E_\theta[\hat{\Theta}_n]$ e $\text{var}_\theta(\hat{\Theta}_n)$, rispettivamente, e sono entrambe funzioni numeriche di θ , ma per semplicità, quando il contesto è chiaro non mostriamo questa dipendenza (da θ).

Abbiamo una terminologia per gli stimatori. Sia $\hat{\Theta}_n$ uno stimatore del valore sconosciuto θ , cioè, una funzione di n osservazioni X_1, X_2, \dots, X_n la cui distribuzione dipende da θ .

- L'errore della stima (estimation error, in inglese), denotato con $\tilde{\Theta}_n$, è definito come $\tilde{\Theta}_n = \hat{\Theta}_n - \theta$.
- Il *bias* di uno stimatore, denotato con $b_\theta(\hat{\Theta}_n)$, è il valore atteso dell'errore della stima: $b_\theta(\hat{\Theta}_n) = E_\theta[\hat{\Theta}_n] - \theta$.
- Diciamo che $\hat{\Theta}_n$ è *unbiased* se $E_\theta[\hat{\Theta}_n] = \theta$, per ciascun valore di θ .
- Diciamo che $\hat{\Theta}_n$ è *assintoticamente unbiased* se $\lim_{n \rightarrow \infty} E_\theta[\hat{\Theta}_n] = \theta$, per ciascun valore di θ .
- Diciamo che $\hat{\Theta}_n$ è *consistente* se la sequenza $\hat{\Theta}_n$ converge al vero valore di θ , in probabilità, per ciascun valore di θ .

Definizione 3 (Massima verosomiglianza). *Sia $X = (X_1, X_2, \dots, X_n)$ un vettore di osservazioni descritto dalla PMF congiunta $p_X(x; \theta)$ la cui forma dipende da θ . Supponiamo di osservare un valore particolare $x = (x_1, x_2, \dots, x_n)$ di X . Allora, una stima di massima verosomiglianza (maximum likelihood, in inglese) è un valore del parametro che massimizza la funzione numerica $p_X(x_1, \dots, x_n; \theta)$*

su tutti i valori di θ :

$$\hat{\theta}_n = \arg \max_{\theta} p_X(x_1, \dots, x_n; \theta).$$



Se X è continua, allora la PMF congiunta $p_X(x; \theta)$ è sostituita dalla PDF congiunta $f_X(x; \theta)$, tale che:

$$\hat{\theta}_n = \arg \max_{\theta} f_X(x_1, \dots, x_n; \theta).$$

Diciamo che $p_X(x; \theta)$ (oppure $f_X(x; \theta)$, se X è continua) è la *funzione di verosomiglianza*.

In molte applicazioni, le osservazioni X_i si assumono indipendenti, e, in tal caso, la funzione di verosomiglianza ha la forma:


$$p_X(x_1, \dots, x_n; \theta) = \prod_{i=1}^n p_{X_i}(x_i; \theta),$$

se X_i sono discrete. In questo caso, è spesso analiticamente oppure computazionalmente conveniente massimizzare il suo logaritmo, cioè la *funzione della log-verosomiglianza*:

$$\log p_X(x_1, \dots, x_n; \theta) = \log \prod_{i=1}^n p_{X_i}(x_i; \theta) = \sum_{i=1}^n \log p_{X_i}(x_i; \theta),$$

su θ . Se X_i sono continue, un ragionamento simile viene applicato.

Il termine "verosomiglianza" deve essere propriamente interpretato. In particolare, avendo il valore osservato x di X , $p_X(x; \theta)$ non è la probabilità che il parametro sconosciuto sia uguale a θ , però è la probabilità che x venga osservato quando il parametro è uguale a θ . Dunque, nel massimizzare la verosomiglianza, ci domandiamo: "Qual'è il valore di θ tale per cui le osservazioni viste hanno più probabilità di risultare?"

Esempio 3. Alice modella il tempo che lei investe ciascuna settimana facendo i compiti come una variabile aleatoria esponenziale con parametro sconosciuto θ . I tempi dedicati per ciascuna settimana sono indipendenti. Dopo aver speso 10, 14, 18, 8 e 20 ore nelle prime 5 settimane, qual'è il suo stimatore di massima verosomiglianza di θ ? 

Esempio 4. Si consideri una sequenza di lanci indipendenti di una moneta, e sia θ la probabilità che esca testa in ciascun lancio. Fissare qualche $n \in \mathbb{N}$ e sia K il numero di teste osservate negli n lanci. Trovare lo stimatore di massima verosomiglianza di θ basandoci su K . \diamond

Esempio 5. Delle particelle instabili vengono emesse da una sorgente e decadono a distanza X , che è esponenzialmente distribuita con parametro sconosciuto θ . Un dispositivo speciale è usato per individuare i primi n eventi delle decadute che occorrono in un intervallo $[m_1, m_2]$. Supponiamo che questi eventi siano registrati a distanze $X = (X_1, X_2, \dots, X_n)$. Dare la forma della funzione di verosomiglianza e della sua versione logaritmica.

◇

Adesso discutiamo il problema semplice, ma importante, nel stimare la media e la varianza di una distribuzione di probabilità. Supponiamo che le osservazioni X_1, \dots, X_n siano i.i.d. con media sconosciuta θ . Lo stimatore più naturale di θ è la media campionaria:

$$M_n = \frac{X_1 + \dots + X_n}{n}.$$

Lo stimatore è non biased, siccome $E_\theta[M_n] = E_\theta[X] = \theta$. Il suo errore quadratico medio è uguale alla sua varianza, che è v/n , dove v è la varianza comune delle X_i ; si noti che non dipende da θ . Inoltre, dalla legge "debole" dei grandi numeri, questo stimatore converge in probabilità a θ , ed è dunque consistente.

Si supponga che oltre allo stimatore della media θ :

$$M_n = \frac{X_1 + \dots + X_n}{n}.$$

siamo interessati anche nello stimare la varianza v . Una scelta naturale è:

$$\bar{S}_n^2 = \frac{1}{n} \cdot \sum_{i=1}^n (X_i - M_n)^2,$$

che coincide con lo stimatore di massima verosimiglianza sotto l'ipotesi della normalità. Usando i fatti:

$$E_{(\theta,v)}[M_n] = \theta, \quad E_{(\theta,v)}[X_i^2] = \theta^2 + v, \quad E_{(\theta,v)}[M_n^2] = \theta^2 + \frac{v}{n},$$

abbiamo che:

$$\begin{aligned} E_{(\theta,v)}[\bar{S}_n^2] &= \frac{1}{n} \cdot E_{(\theta,v)} \left[\sum_{i=1}^n X_i^2 + 2M_n \cdot \sum_{i=1}^n X_i + nM_n^2 \right] \\ &= E_{(\theta,v)} \left[\frac{1}{n} \cdot \sum_{i=1}^n X_i^2 - 2M_n^2 + M_n^2 \right] \\ &= E_{(\theta,v)} \left[\frac{1}{n} \sum_{i=1}^n X_i^2 - M_n^2 \right] \\ &= \theta^2 + v - \left(\theta^2 - \frac{v}{n} \right) \\ &= \frac{n-1}{n} \cdot v. \end{aligned}$$

Dunque \bar{S}_n^2 non è uno stimatore non biased per v , anche se è asintoticamente non biased. Possiamo ottenere uno stimatore non biased per la varianza:

$$\hat{S}_n^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - M_n)^2 = \frac{n}{n-1} \bar{S}_n^2.$$

Il calcolo precedente ci mostra che $E_{(\theta,v)}[\hat{S}_n^2] = v$, cioè che \hat{S}_n^2 è uno stimatore non biased per v , per tutte le n . Tuttavia, per n ragionevolmente grande, gli stimatori \hat{S}_n^2 e \bar{S}_n^2 sono essenzialmente lo stesso.

Esempio 6. Una sorgente emette un numero aleatorio di fotoni K ciascuna volta che è innescata. Assumiamo che la PMF di K sia:

$$p_K(k; \theta) = c(\theta) \cdot e^{-\theta \cdot k},$$

per $k = 0, 1, 2, \dots$, dove θ è la temperatura inversa della sorgente e $c(\theta)$ è un fattore di normalizzazione. Assumiamo anche che l'emissione di ciascun fotone sia indipendente. Vogliamo stimare la temperatura della sorgente innescando ripetutamente e contando il numero di fotoni emessi.

- i) Determinare $c(\theta)$.
- ii) Trovare $E[K]$ e $var(K)$ del numero K di fotoni emessi se la sorgente è innescata solo una volta.
- iii) Derivare l'estimatore di massima verosomiglianza logaritmica, sulla base di K_1, K_2, \dots, K_n , cioè il numero di fotono emessi dalla sorgente innescata n volte.

◇

